Lecture 6—The Finale of EDF 9770: Caveats and Next Steps

- Topics:
  - Summary of what we've covered as "The GLM"
    - Review of steps in GLM analysis
    - Understanding GLM assumptions
  - What to do given untenable GLM assumptions: Repair or Eject?
  - Next steps: 4 Lego building blocks of quantitative methods
    - 1. Linear models (done), 2. likelihood estimation, and 3. link functions
    - 4. Random effects / latent variables → preview of MLM in Fall 2026!

Two Reasons Why You WERE Here

1. "This class fulfills a requirement" (and I just need to pass it).
   - I get it—but I (still) hope to have convinced you otherwise!

2. "I want to learn more about data analysis using **quantitative methods**"!
   - **Quantitative methods = Quantitative data + application of statistical models to answer questions**
   - As I promised, the hard part is not the math—it's the working memory load needed to link language (terminology, notation, code) to logic (data, questions, variables, and models)
   - An important component to doing quantitative methods well is recognizing **when the tools you have will not be sufficient** for the data at hand…
   - First, let's review where we've been this this semester… then where things can go wrong… and a preview of what can be done instead!

Review: Steps in a GLM Analysis

1. Understand the research questions to be answered and the types of variables to be used in answering them

   - This will dictate which variables are involved, and whether they are to be considered predictors or outcomes
     - **Predictor** → explainer: *regressor*, *independent variable* (that you care about specifically or that is manipulated), *covariate* (that someone else cares about or is quantitative)
     - **Outcome** → to be explained: *response variable, dependent variable*, *criterion*

   - Primary types of variables: Quantitative or categorical
     - **Quantitative** → numbers are **numbers** (but may have boundaries)
     - **Categorical** → numbers are **labels** (finite and discrete set of possibilities)
     - *Note that the GLM is for predicting quantitative outcomes only!*

Review: Steps in a GLM Analysis

3. Determine how to include **predictor variables** in models

- **<u>Quantitative</u> variables** should be **centered** by adding or subtracting a constant as needed so that 0 is a meaningful reference point
  - Why? To create a useful fixed intercept at a minimum; also for useful "main effect" slopes of predictors that are part of interaction terms (conditional on moderator=0)
  - But predicted outcomes and model R^2 do not depend on centering…
    (so there are no wrong choices for centering constants, only weird)

- Then consider their **type of relation** with the outcome
  - **Linear** is default, but linear only may not always be plausible…
  - **Quadratic** (by adding predictor^2) allows slope to change directions
  - **Exponential-ish** (through a linear slope of log-transformed predictor) creates a slope that slows down (i.e., capturing diminishing returns)
  - **Piecewise** (linear spline) allows slope to differ across predictor regions
  - This is an empirical question—outcome means per value can help you decide!

Review: Steps in a GLM Analysis

3. Determine how to include **predictor variables** in models

- **<u>Categorical</u> predictors** (numbers are just **labels**) should only be included as-is if they are already binary (0 or 1)
- Otherwise, they need to be represented by $C - 1$ new dichotomous predictors for C categories *(*can be done for you by using a "factor" variable)*

Indicator Coding*
(useful for nominal)

original	New Predictors		
group	AvB	AvC	AvD
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

Sequential Coding
(more useful for ordinal)

original	New Predictors		
group	AvB	BvC	CvD
A	0	0	0
B	1	0	0
C	1	1	0
D	1	1	1

Indicator Coding Done for You: Factor Variables

- Designate a predictor as "**categorical**" in data or syntax → **factor variable in R**
- For a predictor with C categories, the program automatically then creates C new indicator-coded binary predictors, for example "group" with $C = 4$:

New Internal Predictors Mean This:	**A**	**B**	**C**	**D**
IsA	**1**	0	0	0
IsB	0	**1**	0	0
IsC	0	0	**1**	0
IsD	0	0	0	**1**

```
# These two models are equivalent
ModelManual = lm(data=MyData, formula=outcome~1+AvB+AvC+AvD)
ModelFactor = lm(data=MyData, formula=outcome~1+group)
```

original	**New Predictors**		
group	**AvB**	**AvC**	**AvD**
A	0	0	0
B	**1**	0	0
C	0	**1**	0
D	0	0	**1**

- If you enter "group" as a predictor, R then figures out how many of these internal predictor variables are needed—if using an intercept (the default), then it's $C - 1$, not C
- It enters them until it hits that criterion—if it leaves the first one out, then the first category becomes your reference (btw, last is left out instead in SPSS and SAS)

Example: Factor Variables in R

```
# Frequencies per category
table(HSB$program)
```

```
vocation  general academic
     147      145      308
```

Here is a useful tutorial on how to work with factor variables

```
# Reorder levels from default alphabetical to desired order
HSB$program = factor(HSB$program, levels = c("general","vocation","academic"))
levels(HSB$program)
```

```
"general"  "vocation" "academic"
```

```
# Other options for re-ordering using mean of another variable
HSB$program = reorder(HSB$program, HSB$locus)  # Ascending by locus
HSB$program = reorder(HSB$program, -HSB$locus) # Descending by locus

# Assign sequential contrasts using MASS package
# Intercept is grand mean forcing equal size per category
contrasts(HSB$motiv) = contr.sdif(nlevels(HSB$motiv))
contrasts(HSB$motiv)
```

```
  2-1  3-2   4-3
1 -0.75 -0.5 -0.25
2  0.25 -0.5 -0.25
3  0.25  0.5 -0.25
4  0.25  0.5  0.75
```

```
# Convert string (text) into factor variable
HSB$ses = as.factor(HSB$ses_string)
# Convert while renaming and/or re-ordering
HSB$ses = factor(HSB$ses_num, levels = c("1","2","3"),
                 labels = c("low","middle","high"))
```

Factor variables can be convenient, but their lack of transparency can lead to misinterpretation, especially when used in interaction terms…

Example: Factor Variables in R

```
Model1 = lm(data=HSB, formula=math~1+program+career)
obj=LMsummary(Model1)  # Custom output

# Get pairwise comparisons of program
emmeans(Model1, pairwise~program, adjust="none")
# Get joint F-tests using Type III SS
car::Anova(Model1, type=3)
# Wrong version using Type I SS (credit in order)
anova(Model1)
```

```
$emmeans
 program    emmean       SE  df lower.CL upper.CL
 general  48.9923 0.763314 581  47.4931  50.4915
 vocation 46.4455 0.741027 581  44.9901  47.9009
 academic 55.0846 0.623207 581  53.8606  56.3086

Results are averaged over the levels of: career
Confidence level used: 0.95

$contrasts
 contrast           estimate       SE  df t.ratio p.value
 general - vocation  2.54681 1.003126 581   2.539  0.0114
 general - academic -6.09224 0.878352 581  -6.936 <0.0001
 vocation - academic -8.63905 0.908629 581  -9.508 <0.0001
```

Estimated means "marginalizing" over career equally weight all categories!

```
Fixed Effects Table
                    Est    SE      t      p    LCI    UCI
Intercept         49.435 1.382 35.774 <0.001 46.721 52.150
programvocation   -2.547 1.003 -2.539 0.011 -4.517 -0.577
programacademic    6.092 0.878  6.936 <0.001  4.367  7.817
careercraftsman   -1.282 1.792 -0.716 0.475 -4.801  2.237
careerfarmer       3.301 2.791  1.183 0.237 -2.180  8.783
careerhomemaker   -2.031 1.889 -1.075 0.283 -5.742  1.680
careerlaborer     -1.052 2.688 -0.391 0.696 -6.332  4.228
careermanager     -1.032 2.111 -0.489 0.625 -5.178  3.113
careermilitary    -3.482 2.257 -1.543 0.123 -7.915  0.951
careeroperative   -3.347 2.074 -1.613 0.107 -7.421  0.727
careerprof1       -0.033 1.377 -0.024 0.981 -2.737  2.671
careerprof2        1.904 1.511  1.259 0.208 -1.065  4.872
careerproprietor  -2.029 2.155 -0.942 0.347 -6.262  2.203
careerprotective  -2.221 3.025 -0.734 0.463 -8.162  3.719
careersales       -0.344 2.704 -0.127 0.899 -5.654  4.966
careerschool       3.661 2.346  1.561 0.119 -0.946  8.268
careerservice     -2.767 1.959 -1.412 0.158 -6.614  1.081
careertechnical    3.473 1.848  1.880 0.061 -0.156  7.103
careernot working -0.251 3.022 -0.083 0.934 -6.187  5.685
```

```
Anova Table (Type III tests)

Response: math
             Sum Sq  Df  F value Pr(>F)
(Intercept)  89116   1 1279.759 <2e-16
program       7237   2   51.962 <2e-16
career        2121  16    1.903 0.0178
Residuals    40458 581
```

Car::Anova provides the same F-tests as my R2compare custom function, but without needing to fit a nested model (but without effect sizes)

Reviewing the Steps in a GLM Analysis

4. Get what else you need that **isn't directly given**, like:
   - **Predicted outcomes** (e.g., for non-reference groups)
   - t-tests (numerator DF=1) of **new fixed single slopes** created using linear combinations of fixed intercept and slopes
     - e.g., other predictor category mean differences (such as category B v D)
     - e.g., differences between fixed slopes (such as between sequential slopes between categories or between predictor slopes on same scale)
     - e.g., conditional slopes for predictors in interaction terms at other moderator values
   - F-tests (DF>1) for joint tests of **fixed slopes lumped together**
     - e.g., for "omnibus" (= overall) effects of categorical predictors
     - e.g., for testing changes to model R^2 for a set of new slopes (avoid ambiguous order)

Reviewing the Steps in a GLM Analysis

5. Get **effect sizes** (to convey absolute relationship size independent of statistical significance that is also governed by N):

   - Always ok: **per-slope partial r** (aka, "**eta**") **or Cohen's d** → provides size of unique contribution relative to it + residual
     - Useful for predicting power when planning similar analyses
     - Not useful in relation to model R^2 (because not out of total variance)
     - Can be inflated by adding predictors to reduce residual variance
   - Can be ok: **per-slope semi-partial R^2** (aka, "**eta-squared**") → provides amount of model R^2 due to that predictor
     - Slopes of binary-coded predictors to represent a categorical variable need to be lumped together first (because they are not independent)
     - Slopes of predictors involved in interaction terms (including quadratic terms) are conditional on moderators = 0 (so effect size can change based on centering)

Effect Size Definitions by Family and Metric

- My preference is to compute partial r or Cohen's d (which is also *partial*) per fixed slope, but semi-partial R^2 per conceptual predictor
- I usually avoid partial R^2 due to its greater chance of misinterpretation

	Definition using r	**Definition using R^2**
Partial	X–Y correlation after controlling **X and Y** for all other predictors	$\frac{SS_x}{SS_x+SS_{res}}$ = R^2 for variance remaining after other predictors
Semi-partial	X–Y correlation after controlling **only X** for all other predictors	$\frac{SS_x}{SS_{total}}$ = amount of model R^2 due to that predictor's slope(s)

Reviewing the Steps in a GLM Analysis

6. Write it up and turn it in!

- Models to be reported are likely only a subset of all the models estimated—report those that tell the honest story in answering your RQs and include:
  - **Analytic method**: modeling family or approach, software and version (including packages), how predictors were centered or coded (i.e., who the reference is)
    - Equations are useful when done correctly—use the proper notation in the equation editor!
  - **What happened**: in both "stat-ese" and regular language (see my results sections in class materials and in homework assignments for example templates)
    - Per slope: Estimate, standard error, p-value, effect size
    - Per model: F-value, both DF, residual variance, p-value, R^2
    - Always guide the reader—tell them explicitly why they should care (i.e., what RQ is relevant)
  - Consider "**supplemental material**" for any results you don't have room for, as well as equations and syntax—these resources can help you get cited!

Labels for What We Covered This Semester

Intro to **General Linear Models** (GLMs) as a one-stop shop, but *only for predicting one conditionally normal outcome per person in a single dimension of sampling*

- **Quantitative** predictors = *"(linear) regression"*
  - 1 numeric predictor variable = *"simple (linear) regression"*
  - 2+ numeric predictor variables = *"multiple (linear) regression"*
  - Includes linear and nonlinear (e.g., quadratic, exponential-ish) relations
- **Categorical** predictors = *"analysis of variance (ANOVA)"*
  - 1 two-group predictor variable = *"independent-samples t-test"*
  - 1 three-or-more-group predictor variable = *"one-way ANOVA"*
  - 2+ group predictor variables = *"two-way (or factorial) ANOVA"*
- Both kinds of predictors = *"analysis of covariance (ANCOVA)"*
- We covered **moderation** (via interactions) of some kinds, too!

Btw: 2+ way "ANOVA" usually implies all possible interaction terms, even if nonsignificant!

The Missing Step 2: Select Model Family!

- The GLM requires several things to be plausible for the results to be believable—these are called "**model assumptions**"

- Two types of consequences of violated assumptions:
  - On fixed slope **estimates**
    - Wrong estimates → wrong depiction of variable relations
    - Labeled "primary" by Darlington & Hayes (2017)
  - On fixed slope **standard errors**
    - Wrong conclusion about inconsistency → wrong p-value
    - Labeled "secondary" by Darlington & Hayes (2017)

- Some problems can be fixed by modifying the GLM (model format or its estimation), but some can't!
  - There is much confusion over what is actually assumed… so let's discuss:

Is "Linearity" an Assumption?

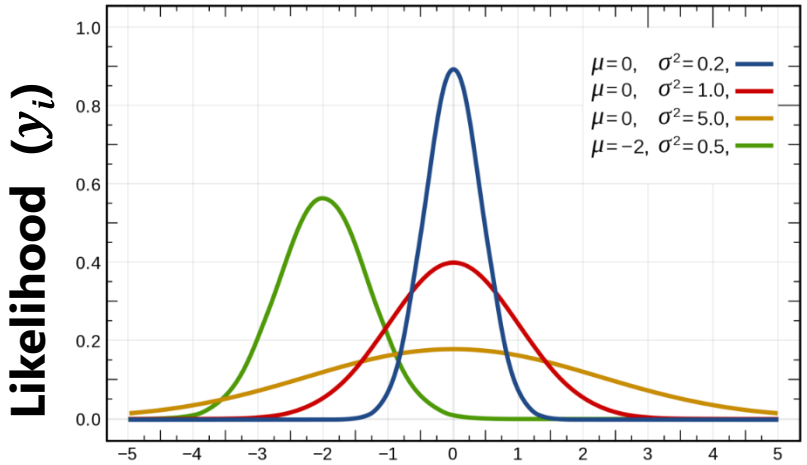
- In a word, **NO**! This is a misconception given that **linear slopes** are the universal default for the effect of quantitative predictors

- We've seen that GLMs can include nonlinear relations of quantitative predictors and outcomes, but the term "**nonlinear**" **needs clarified**:
  - "**Nonlinear in the variables**" (as we've used) means adding predictors that create nonlinear outcome relations in a model of linear form
    - e.g., squared predictors → quadratic form of relation (i.e., parabola happy or sad face)
  - "**Nonlinear in the parameters**" means a model that does not use the "constant*variable + constant*variable" linear form
    - e.g., a truly exponential model: $y_i = \beta_0 + \beta_1\left[e^{\beta_2(x_i)}\right]$

- The GLM (and any model!) a**ssumes the functional form of the predictor relations is correct**, including the potential for both nonlinear relations and interaction slopes
  - Otherwise, your characterization of variable relations may be incorrect
  - Best tested by adding slopes to the model and seeing if they are needed

Is "Normality" an Assumption?

- Of the predictors? Of course not!
  - *Otherwise, ANOVA (with categorical predictors) could not be a thing!*
- Of the marginal (original) outcome? Also NO!
- Instead, the **GLM assumes the e_i residuals**—the leftover, **conditional** outcome—have a **normal distribution**
  - The **normal** distribution describes symmetric, continuous variables
  - Uses two parameters: **mean** (conditional on predictors, given by $\hat{y}_i$) and **1 variance** (σ_e^2 for the residuals)

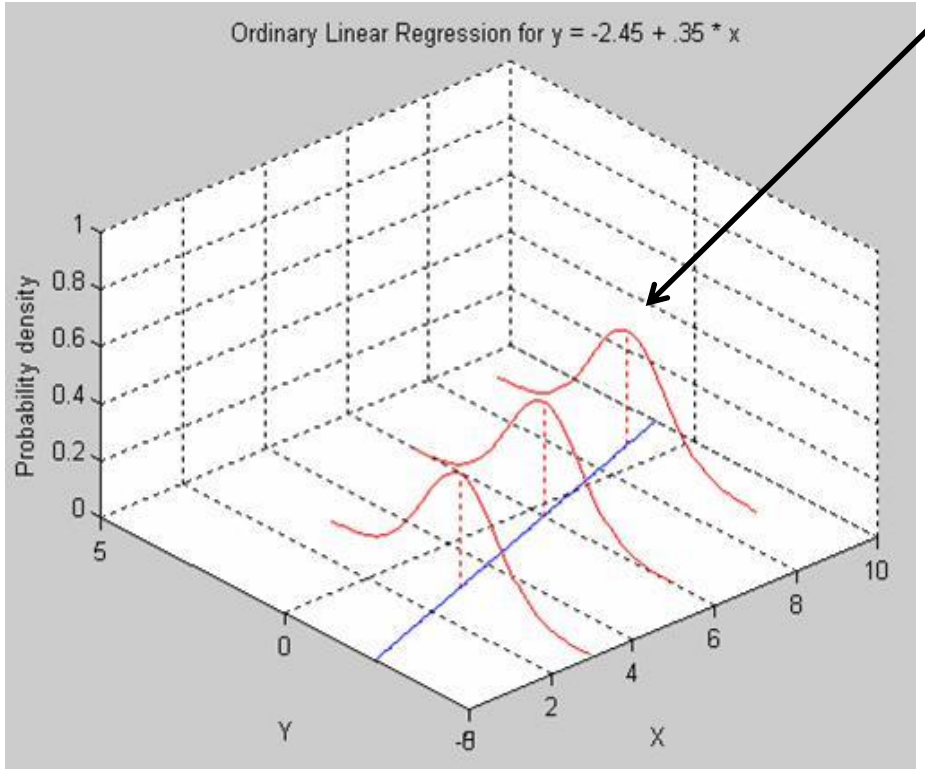


- Stand-alone textbook chapters on "data cleaning" and "data transformation" and "outlier analysis" are really problematic!
  - Because residuals are only possible in the context of a model!

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

Normal Distribution → Constant Variance

- Because GLM residuals should have a normal distribution, this means they should have constant variance—that the same residual variance applies to all cases → "**homoscedasticity**" = "**homogeneity of variance**" = "**constant variance**"



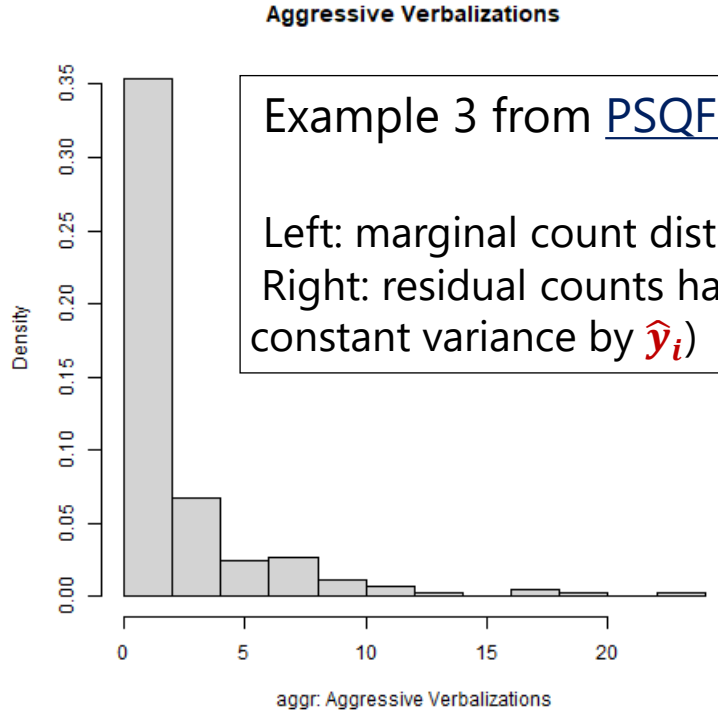
Otherwise, "**heteroscedasticity** " = "**heterogeneity of variance**" → model predicts differentially well across $\boldsymbol{x_i}$ (or $\boldsymbol{\hat{y}_i}$ more generally)

Below: "Not good" → $\boldsymbol{\sigma_e^2}$ increases as the x_i predictor increases (→ fan shape), often due to lower boundary in the scale of the outcome (as in counts)

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

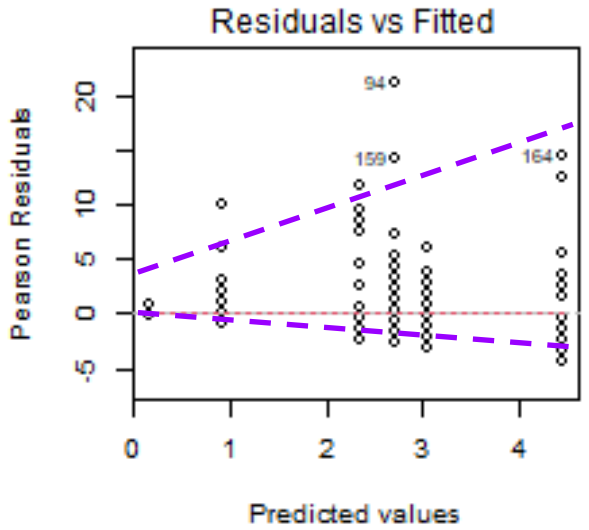
Normal Distribution → Constant Variance

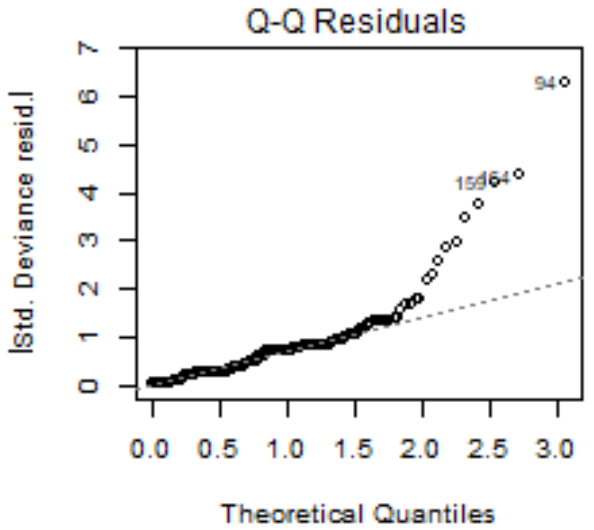
- In practice, both normality and constant variance of the e_i residuals may not hold in quantitative outcomes with boundaries (i.e., continu-ish)
  - e.g., proportion correct outcomes are bounded by both 0 and 1, and **counts** are bounded at 0, so residual variance will shrink as $\hat{y}_i$ approaches these ends



Example 3 from PSQF 6270

Left: marginal count distribution
Right: residual counts have non-constant variance by $\hat{y}_i$)





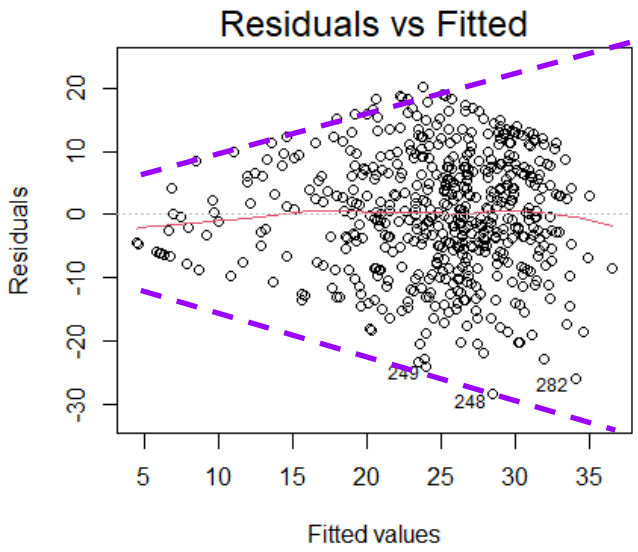
Residual plots in R made with just: `plot(Modelname)`

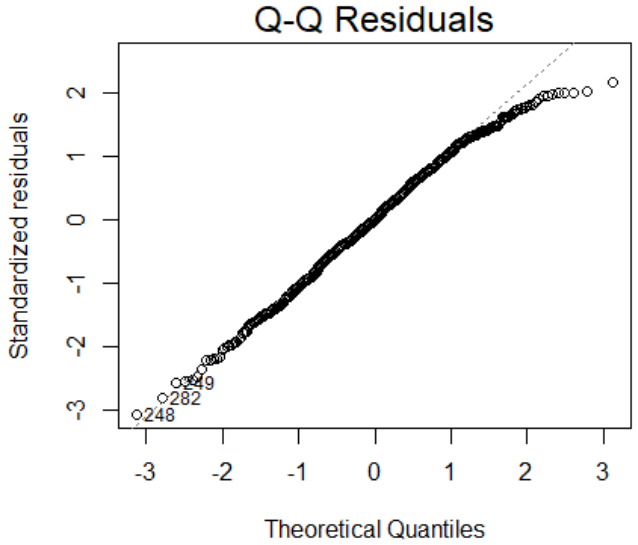
What to Do about Non-Constant Variance

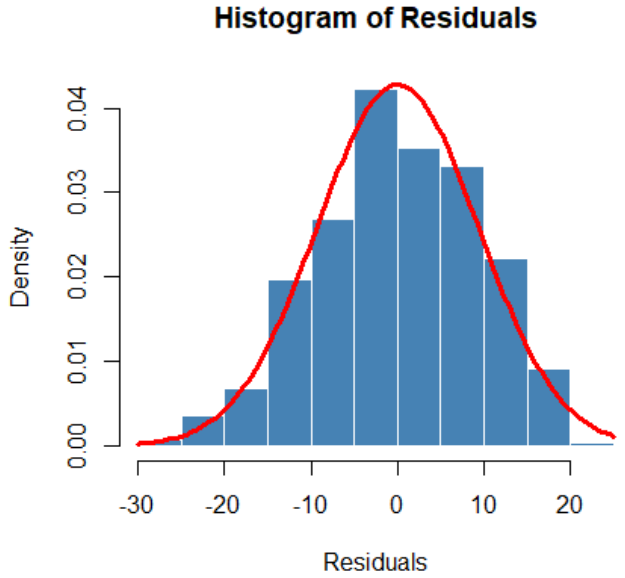
- Residual non-normal residual distributions are not the problem: **Non-constant residual variance** can create **incorrect SEs and p-values**
- For outcomes that are continu-ish, the problem of non-constant variance can be mitigated by **changing the way the SEs are computed**:
  - Request **heterogeneity-consistent** SEs
    - aka, "**sandwich**" estimators: Using matrix notation, predictors are the "bread"; weighted residuals are the "meat" (labeled HC0 to HC4; I will show best-behaving HC3 next)
    - HC3: Divide each e_i^2 by $(1 - leverage_i)^2$ to correct asymptotic covariance matrix of fixed effects (with SE^2 on the diagonal), reducing influence of high-leverage points
    - Lets the residuals tell us directly about the uncertainty of each fixed effect estimate
  - Request bootstrapped SEs (but results will change without setting a seed!)
    - **Bootstrapping**: Sample same N with replacement repeatedly, re-estimate model, get empirical SD of estimates across resamples as new empirical SE

Residuals for Eq 2.9 Model in Example 5

- Example 5 Model 3: cognition outcome was simulated to have a normally-distribution residual $\boldsymbol{e_i}$ with constant variance $\boldsymbol{\sigma_e^2}$
  - Residual plots show plausible normality, but non-constant variance *(because a sex*dementia interaction is still missing relative to the correct population model!)*







Example 5 (Chapter 2) Data: "Robust" SEs

- Cognition outcome was simulated to have a normally-distributed residual with constant variance, so results don't change much using robust SEs (below right)
  - Biggest difference for sex and DemNC slopes (which are mis-specified!)

```
print("Model 3: Remove 2 Sex Interactions, Add Age by Grip Interaction (Equation 2.9)")
Model3 = lm(data=Example5, formula=cognition~1+age85+grip9+sexMW+demNF+demNC+age85:grip9)
obj=LMsummary(Model3)  # Custom output
# Heterogeneity-corrected SEs and CIs using lmtest and sandwich packages
coeftest(x=Model3, vcov=vcovHC(Model3, type="HC3"))
coefci(Model3, vcov=vcovHC(Model3, type="HC3"))
```

Fixed Effects Table

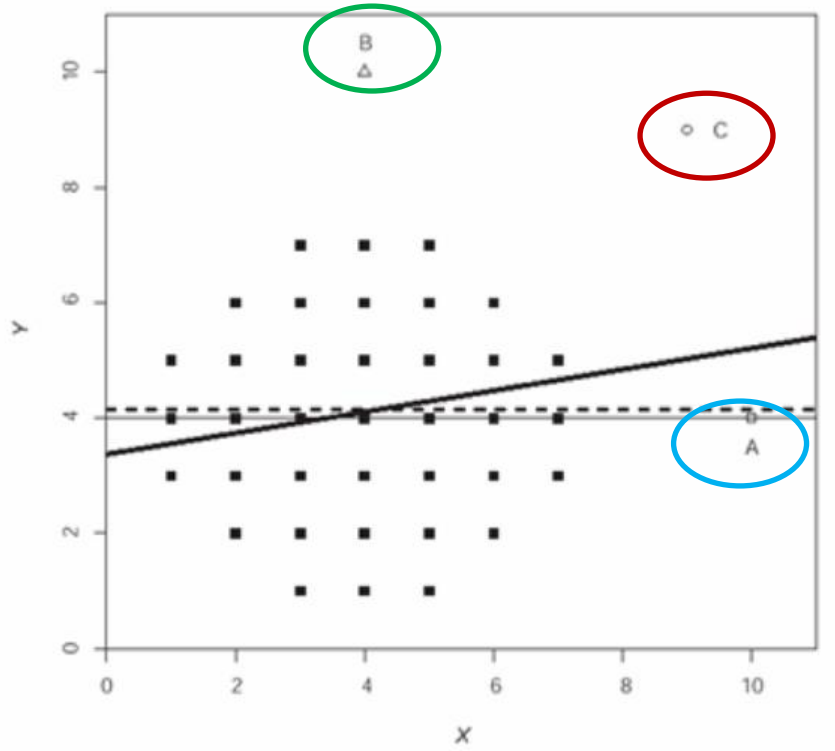
	Est	SE	t	p
Intercept	29.408	0.695	42.319	<0.001
age85	-0.334	0.120	-2.775	0.006
grip9	0.619	0.149	4.164	<0.001
sexMW	-3.456	0.887	-3.895	<0.001
demNF	-5.923	1.014	-5.843	<0.001
demNC	-16.300	1.513	-10.777	<0.001
age85:grip9	0.123	0.041	3.035	0.003

t test of coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.4078	0.7073	41.58	< 2e-16
age85	-0.3340	0.1201	-2.78	0.0056
grip9	0.6194	0.1540	4.02	0.000065725
sexMW	-3.4556	0.9213	-3.75	0.0002
demNF	-5.9225	1.0238	-5.78	0.000000012
demNC	-16.3004	1.1619	-14.03	< 2e-16
age85:grip9	0.1230	0.0424	2.90	0.0038

Problematic Participants

- Skewed residuals can also be due to extreme values (known as "**outliers**"), of 3 kinds:
- "**Distance**" = extreme on y_i (**B**)
- "**Leverage**" = extreme on x_i (**A**)
- "**Influence**" = impact on slope (**C**)
  - Measured by per-person values for:
    - **Cook's distance** = how much $\hat{y}_i$ values would change without that person (is actually "influence", not "distance")
    - **dfBeta** = how much each β fixed slope would change without that person
  - The key is to look for **relatively** high values (absolute cut-offs don't really work in practice)



What to do with any high influence cases? There are no good uniform solutions… it depends on how much you believe the aberrant cases are representative…

Quantile Regression: A Better Way of Addressing Outliers

- If you are concerned about outliers' potentially large influence on the GLM solution, a useful alternative is **quantile regression**
  - Rather than predicting the conditional mean as in GLMs, in quantile regression the conditional median (50th percentile) is predicted instead
  - Robust to outliers AND you can choose *any* percentile, not just 50th!
  - Maximizes the sum of the absolute value of residuals rather than squared residuals
  - Requires resampling (e.g., bootstrapping) to get standard errors and p-values

- Analogous to **predictor by outcome-level interactions**—the effect of predictors may differ at different points along the quantitative outcome
  - e.g., Does a student intervention help low-performing students more than it helps high-performing students?
  - e.g., In older adults, does age predict response time to a greater extent among slower responders than among faster responders?

When to Eject from the GLM to a New Model Family

- For some outcome variable formats, the assumptions of normally-distributed residuals with constant variance will **never** be plausible (e.g., categorical outcomes)
  - Then you need a **general*ized* linear model** instead of a general linear model (where "ized" → not normal) involving link functions and alternative distributions

- Any linear model using observed variables assumes they are measured with perfect reliability—highly unlikely in social sciences with "squishy" constructs
  - But unmodeled measurement error can reduce the variable relations while also creating too-small SEs, which is why we need **latent variables** (built from measurement models)

- The most important GLM assumption is that the $\boldsymbol{e_i}$ **residuals are independent**—that all the reasons why any pair of y_i outcomes would be more related than others are already built into the model as fixed effects of predictor variables
  - Simplest example violation—pre-test and post-test for same people
  - More generally, correlated ("**dependent**") residuals result from sampling over multiple dimensions simultaneously (e.g., multiple students from multiple schools, multiple occasions from multiple persons), which is why we need **correlated residuals and/or random effects**

Quantitative Methods: A Lego-Inspired World View

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

My goal today:

a) describe these **4 Legos**

b) use them to provide the "big picture" of future coursework or self-study

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, and now at Clemson: Multivariate Modeling (EDF 9780)
  - In general: A Lego-based **philosophy** for learning quantitative methods (developed in cahoots with Jonathan Templin)

The 4 Lego Building Blocks

The Legos covered in the multivariate "bridge" 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)—to be offered Fall 2026 as EDF 9850

   (b) for modeling relations of **"unobserved constructs"** using confirmatory factor analysis (CFA), item response theory (IRT), and structural equation models (SEM)—to be offered Spring 2027 as EDF 9840

How the Lego Blocks Fit Together

1. **Linear models are the mechanism** by which most research questions will be answered, 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**
  - **(Ordinary) least squares ("OLS")** is taught first, but is greatly limited in practice
  - **Maximum likelihood ("ML")** picks up where least squares leaves off
  - **Bayesian** picks up where ML gives up (by adding prior distributions)

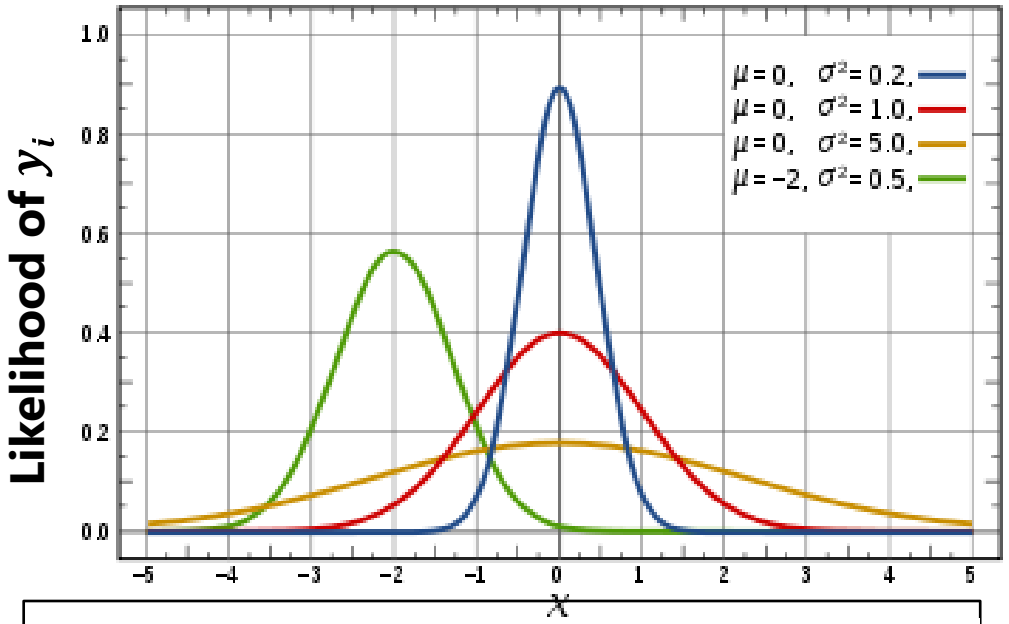
2. Estimation via Maximum Likelihood

- Ordinary Least Squares (OLS) can find answers in **some** kinds of data
  - "Best" fixed effects are those that **minimize the sum of squared residuals**
  - 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 actually computational shortcuts to REML

- **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:

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

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:
  - predicted value $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ → $\boldsymbol{\mu}$
  - Residual variance $\boldsymbol{\sigma_e^2}$ → $\boldsymbol{\sigma^2}$
- 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!

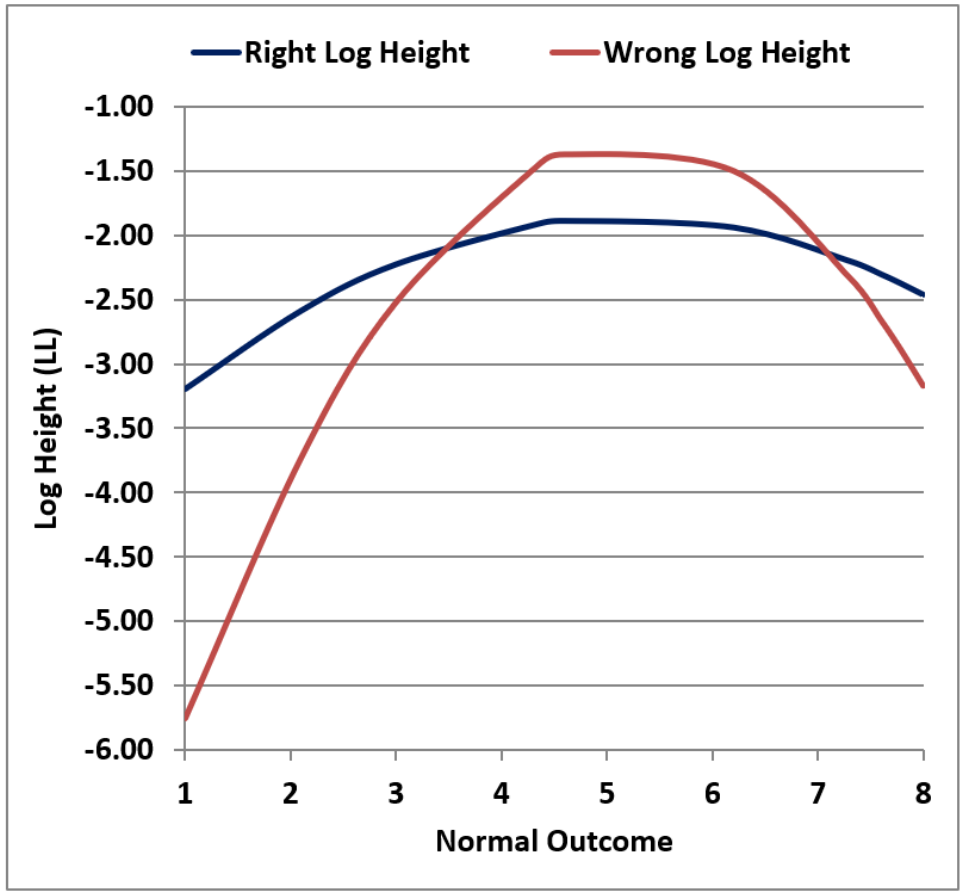
Image borrowed from: https://en.wikipedia.org/wiki/Normal_distribution

ML via Excel NORMDIST: Example

Key idea: Normal Distribution formula → data height
Right answers = tallest possible function for everyone!

	Right Values	Wrong Values
Mean	5.19	5.24
Variance	6.56	2.00
ML Variance	*5.90*	*1.80*

Normal Outcome	Right Height	Right Log Height	Wrong Height	Wrong Log Height
1.0	0.04	-3.20	0.00	-5.76
2.1	0.08	-2.59	0.02	-3.73
3.0	0.11	-2.22	0.08	-2.52
4.3	0.15	-1.92	0.23	-1.49
4.6	0.15	-1.89	0.25	-1.37
6.2	0.14	-1.94	0.22	-1.50
7.3	0.11	-2.20	0.10	-2.33
7.6	0.10	-2.30	0.07	-2.66
7.8	0.09	-2.38	0.05	-2.90
8.0	0.09	-2.46	0.04	-3.17
SUM = ML LL = taller is better		-23.09		-27.42
ML: -2LL = smaller is better		46.19		54.84



What's so great about "normal"?

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

- **The magic of ML:** if your residuals won't be 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

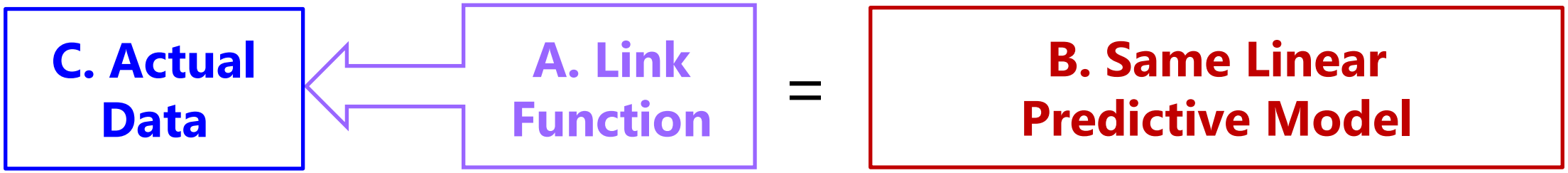
* Note: this is related to traditional levels of measurement, but I am approaching it from more of a "how-to-predict them" perspective

- 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, most 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

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

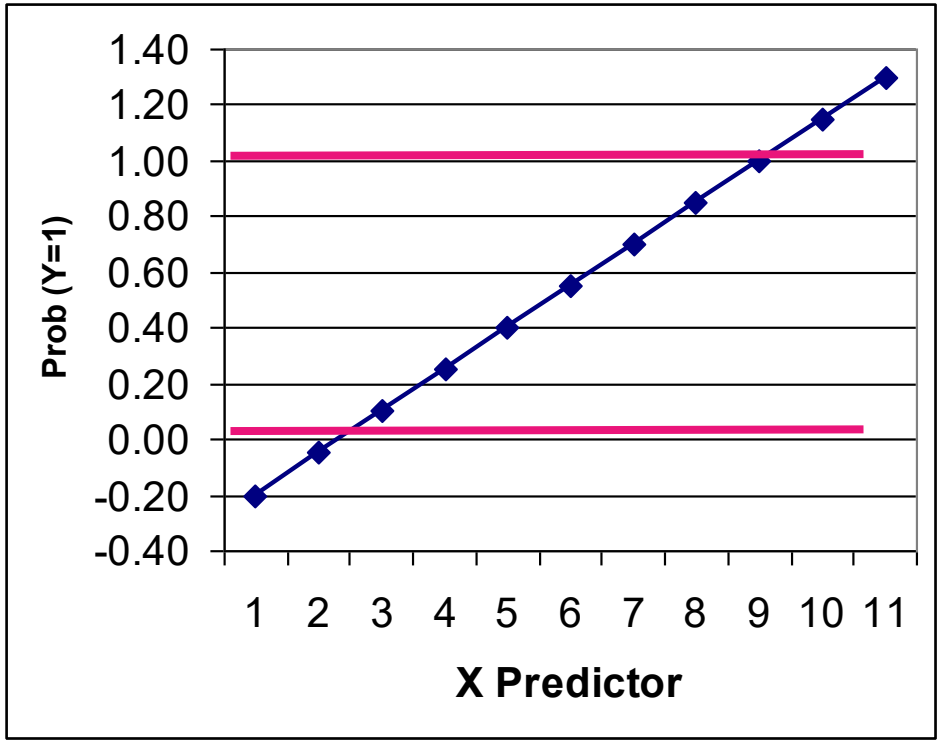
C. **Conditional Distribution**: How the outcome residuals could be distributed given the possible values of the outcome

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...

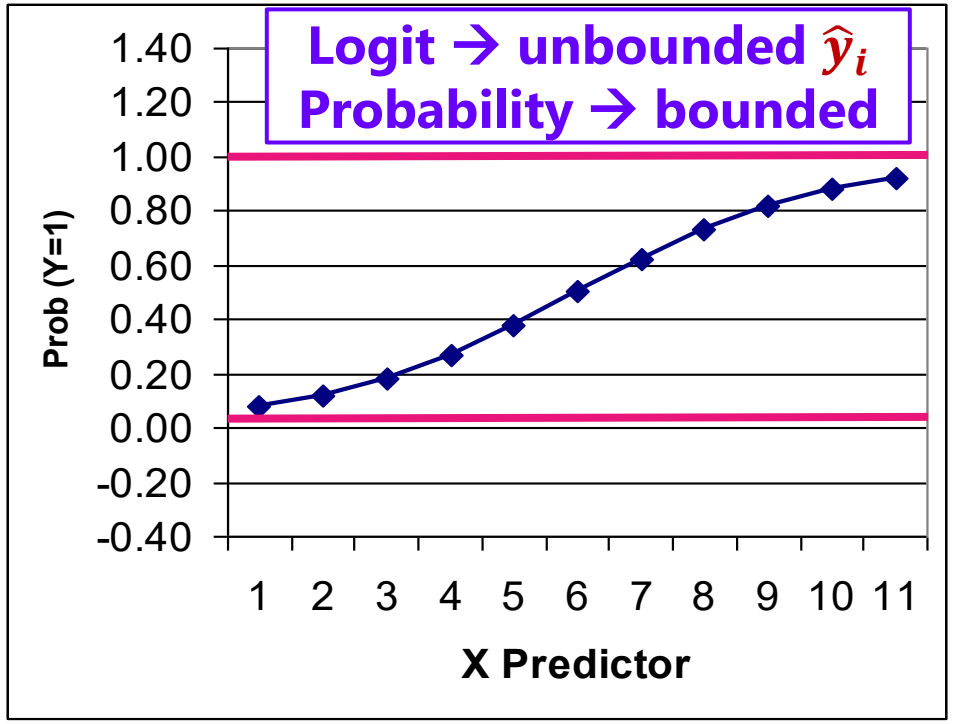
$$p(y_i = 1) = \beta_0 + \beta_1(x_i)$$



To this...

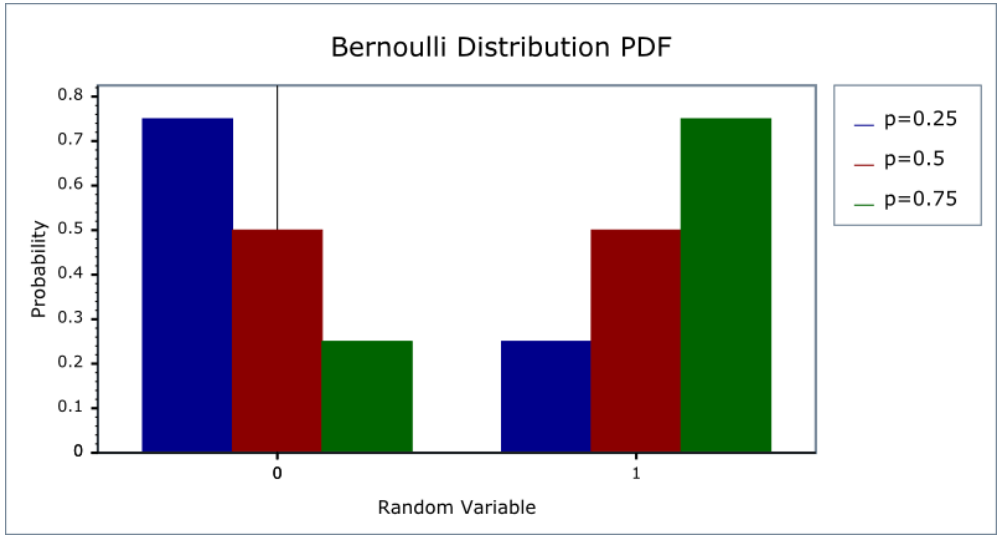
Logit or "log-odds" link

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



Btw, "**probit**" links do the same thing, just on a z-score scale (variance = 1) instead of a logistic scale (variance = $\pi^2/3$)

Bernoulli Distribution: Binary Variables



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)	π_j	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
Variance	σ_j^2	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

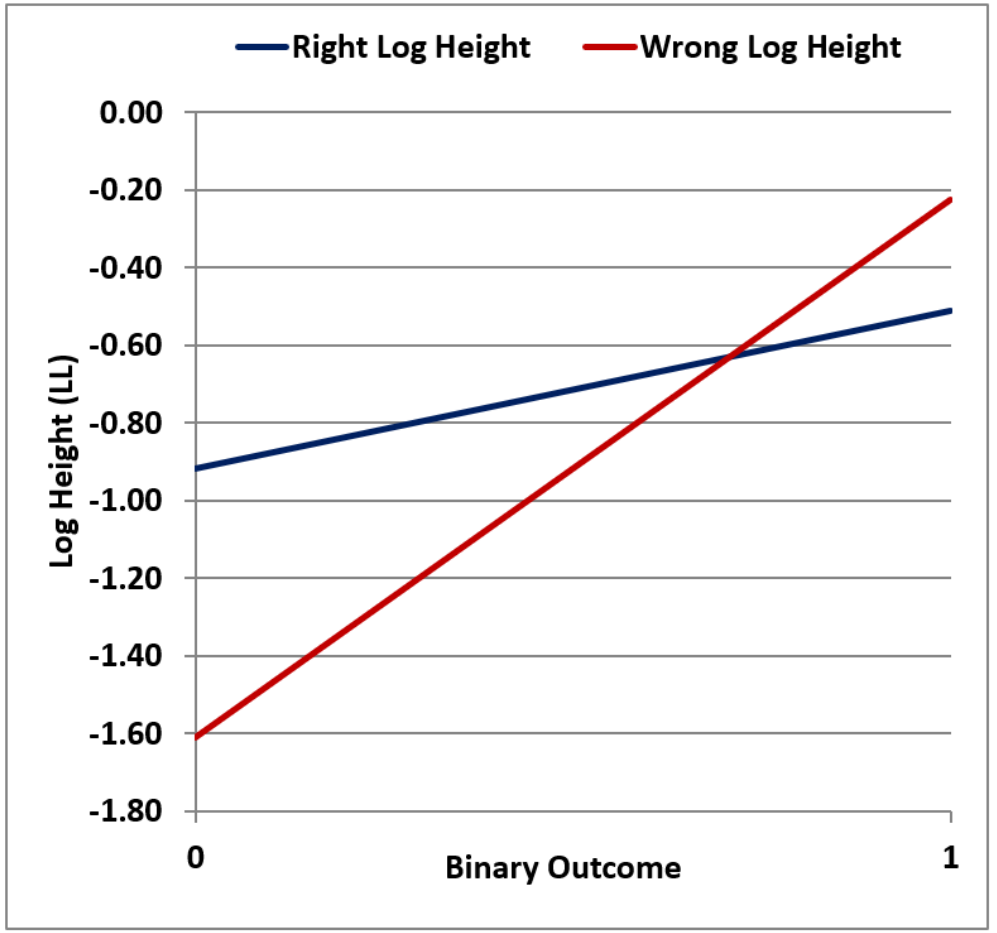
Image borrowed from: https://www.boost.org/doc/libs/1_70_0/libs/math/doc/html/math_toolkit/dist_ref/dists/bernoulli_dist.html

ML via Excel: PDF: $f(y_i) = (p_i)^{y_i}(1-p_i)^{1-y_i}$

Key idea: Bernoulli Distribution formula → data height
Right answers = tallest possible function for everyone!

	Right Values	Wrong Values
Mean	0.60	0.80
Mean in Logits	0.41	1.39

Binary Outcome	Right Height	Right Log Height	Wrong Height	Wrong Log Height
0	0.40	-0.92	0.20	-1.61
0	0.40	-0.92	0.20	-1.61
0	0.40	-0.92	0.20	-1.61
0	0.40	-0.92	0.20	-1.61
1	0.60	-0.51	0.80	-0.22
1	0.60	-0.51	0.80	-0.22
1	0.60	-0.51	0.80	-0.22
1	0.60	-0.51	0.80	-0.22
1	0.60	-0.51	0.80	-0.22
1	0.60	-0.51	0.80	-0.22
SUM = ML LL = taller is better		-6.730		-7.777
ML: -2LL = smaller is better		13.460		15.553



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/Probit** + **Bernoulli**
  - **Ordinal** or **Nominal** → **Logit/Probit** + **Multinomial**
  - **Proportion** → **Logit/Probit** + **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**

Alternative Link Functions for C Categories, Each Built Using $C - 1$ Submodels

- **Cumulative logit/probit (used in IRT "graded response"):**
  ***lower vs. higher category* (using all categories in each submodel)**

  0 vs. **1,2,3** **0,1** vs. **2,3** **0,1,2** vs. **3**

  - **Slopes** usually constrained **equal** across submodels by default → "proportional odds"

- **Adjacent category logit/probit (used in IRT "partial credit"):**
  ***each next highest category* (2 categories per submodel)**

  0 vs. **1** **1** vs. **2** **2** vs. **3**

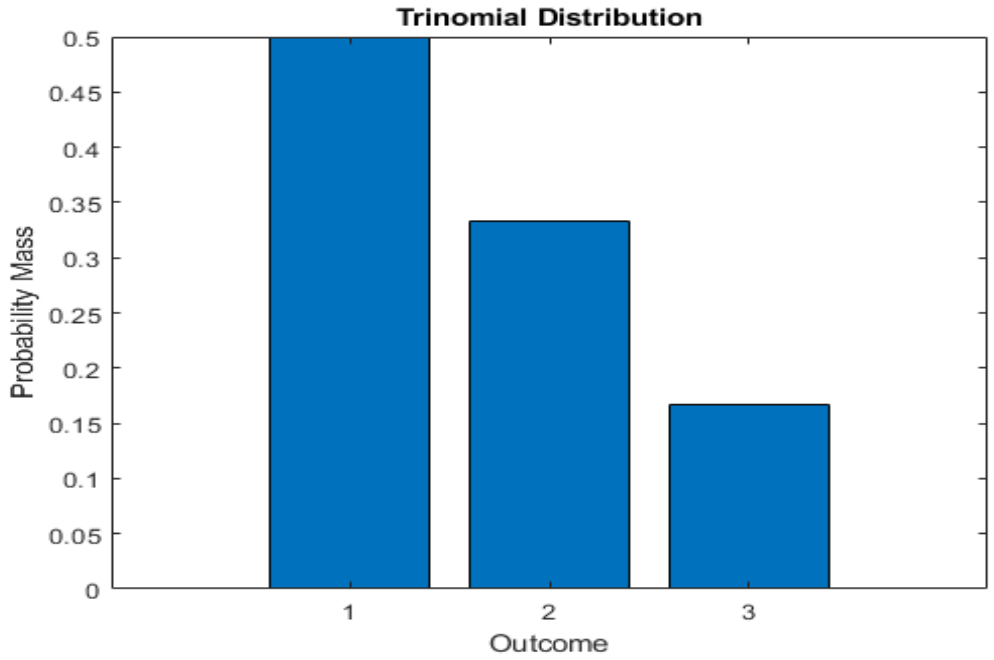
  - **Slopes** usually constrained **equal** across submodels by default

- **Baseline category logit (used in IRT "nominal response"):**
  ***reference (=0 here) vs. each other category* (2 categories per submodel):**

  0 vs. **1** **0** vs. **2** **0** vs. **3**

  - **Slopes usually not constrained equal** across submodels by default

Multinomial Categorical Distribution: Ordinal or Nominal Outcomes



- For example, $C = 3$ possible responses of $c = 1,2,3$, an observed $y_i = c$, and indicators I if $c = y_i$

$$f(y_i = c) = p_{i1}^{I[y_i=1]} p_{i2}^{I[y_i=2]} p_{i3}^{I[y_i=3]}$$

$= p_1(1)$ if 1,
$p_2(1)$ if 2,
$1 - (p_1 + p_2)$ if 3

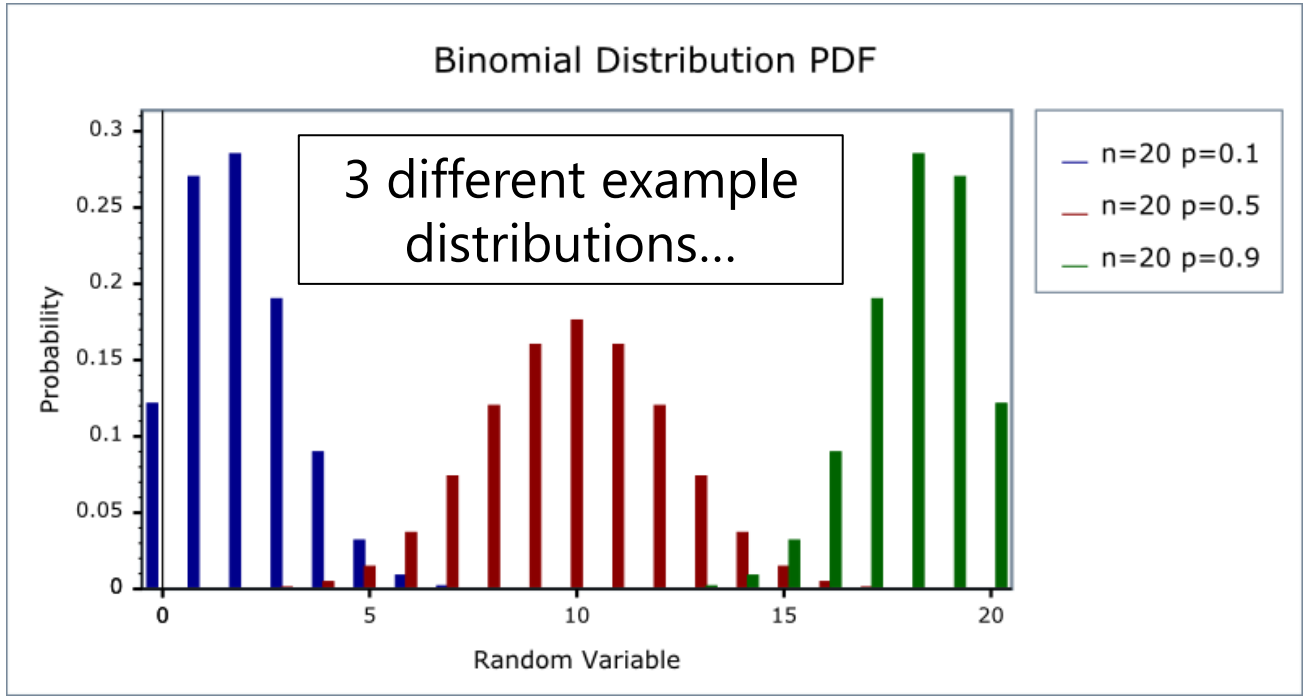
The multinomial distribution has $C - 1$ p mean parameters, called p_c, which create the proportion in each category (so variance is not a separate thing)

Image borrowed from: https://www.mathworks.com/help/stats/work-with-multinomial-probability-distribution-objects.html

Binomial Distribution: Proportions

- The discrete **binomial** distribution can be used to predict $\boldsymbol{c}$ "events" given $\boldsymbol{n}$ trials (**bounded** above and below, so still using a **logit/probit link** function)
  - Bernoulli for binary = special case of binomial when n=1
  - $Prob(y_i = c) = \frac{n!}{c!(n-c)!} p^c (1-p)^{n-c}$

p = probability of 1



Mean = np

Variance = $np(1-p)$

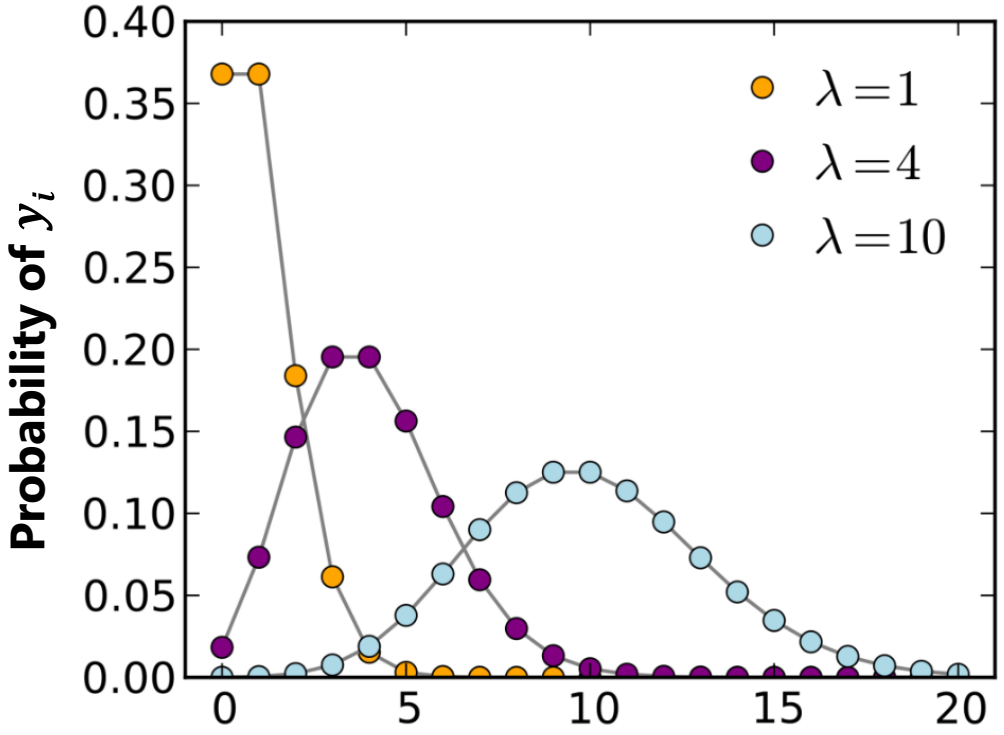
Skewness = $\frac{1-2p}{\sqrt{np(1-p)}}$

Image borrowed from: https://www.boost.org/doc/libs/1_42_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/binomial_dist.html

Negative Binomial (NB) = "Stretchy" Poisson: Counts

Poisson distribution:

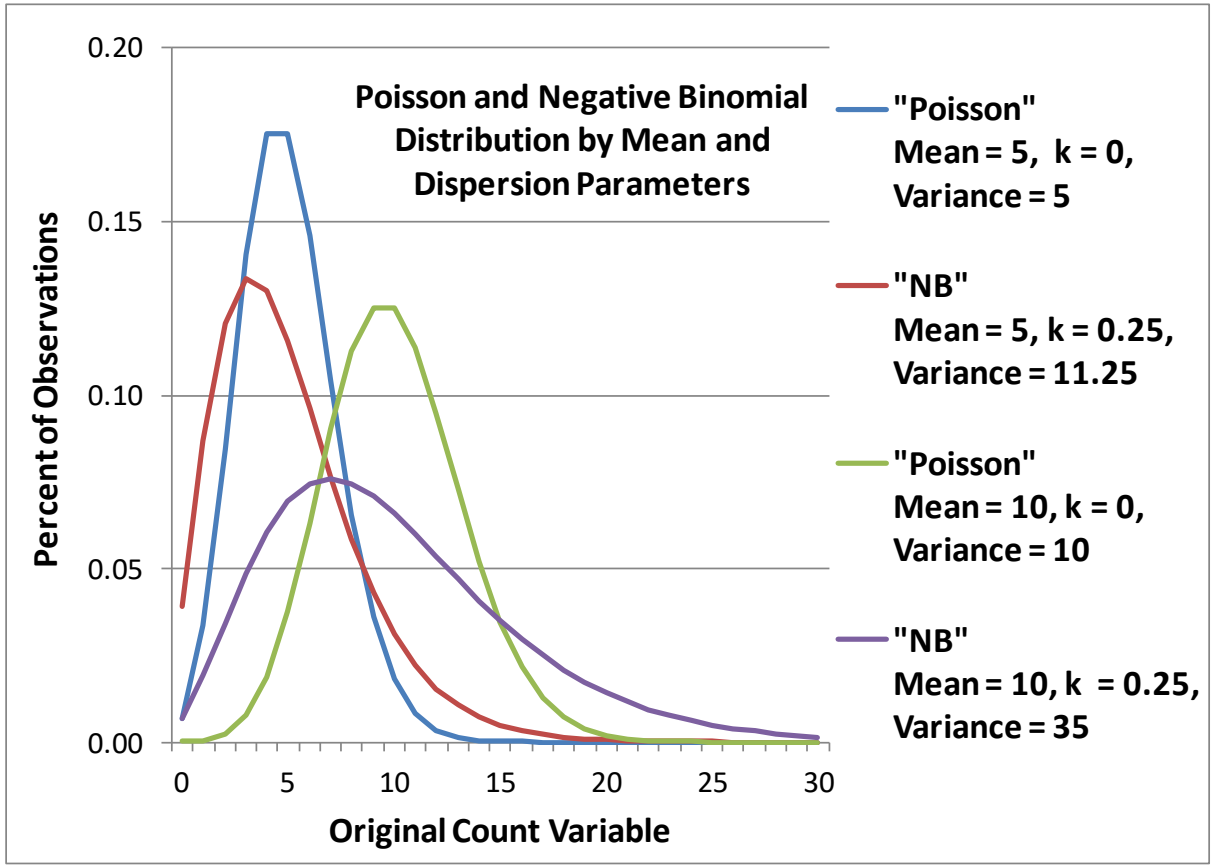
Mean = Variance = λ



Either way, natural log link function keeps predicted counts > 0

Negative binomial distribution:

$\text{Mean} = \lambda, \text{Dispersion} = k, \text{Var}(y_i) = \lambda + k\lambda^2$



Left image borrowed from: https://en.wikipedia.org/wiki/Poisson_distribution

Zero-Inflated Count Variables

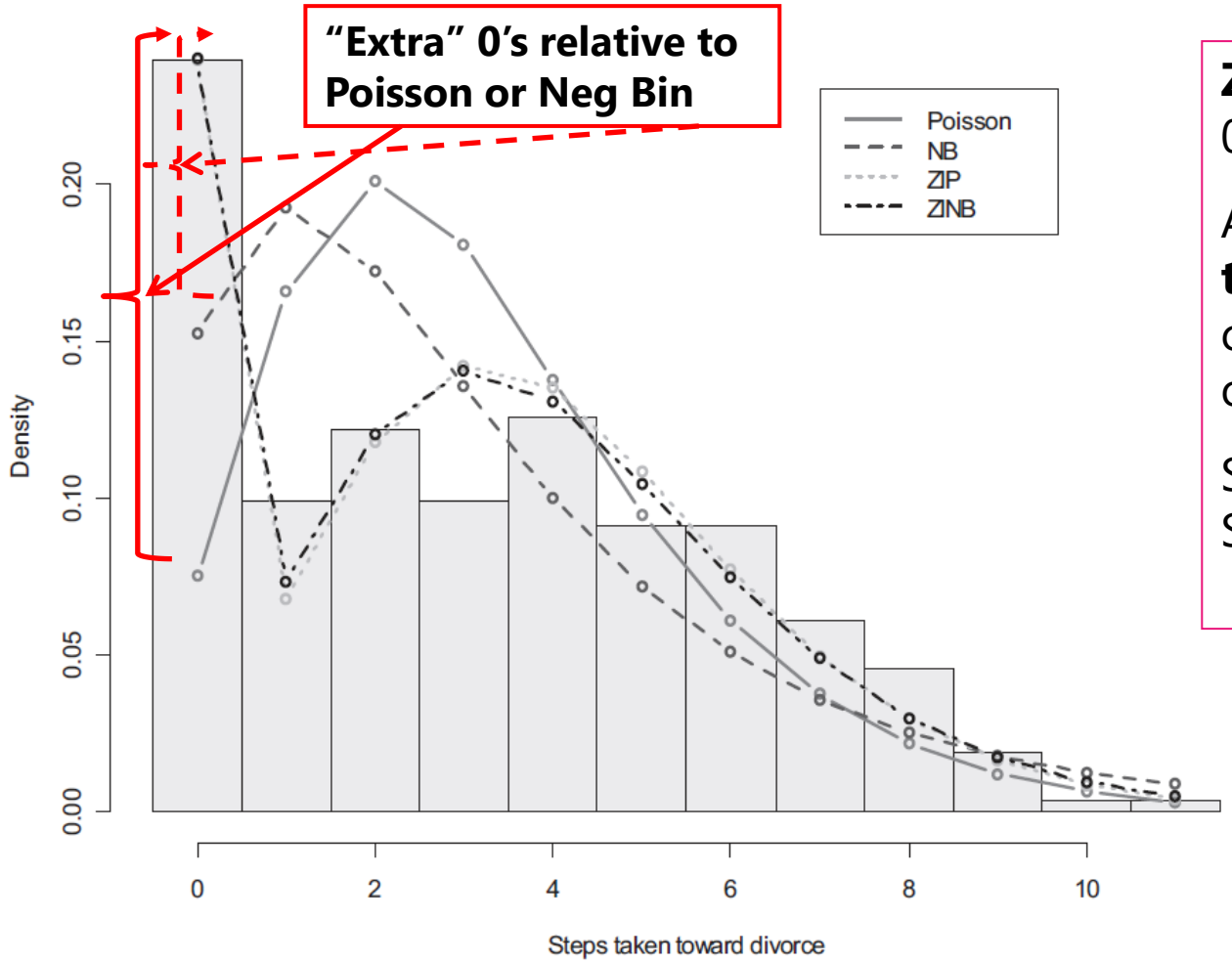


Figure 1. Histogram of Marital Status Inventory with predicted probabilities from regressions. NB = negative binomial; ZIP = zero-inflated Poisson; ZINB = zero-inflated negative binomial.

Zero-inflated count models allow an excess of 0 values beyond expected count distribution

Alternatively, **hurdle** models (for counts) and **two-part** models (for positive continuous) create two submodels for "**if and how much**" outcomes (my own terminology):

Submodel 1: Is y = 0 or not? → Binary outcome
Submodel 2: How much if y > 0? → Quantitative outcome (truncated count or log-normal)

Image borrowed and doctored from: Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*(4), 726-735.

Intermediate Summary: Legos 1, 2, and 3

- To use linear model to predict a non-continuous outcome, we need generalized models: First, add a **link function** (Lego #3) to keep predicted outcomes within possible range (e.g., from 0 to 1 or > 0)
  - Second, use **likelihood estimation** (Lego #2) to select a better-matching distribution than normal (that has separate mean and variance as its parameters) as the formula for height (to find the answers that make the data the tallest)
  - These alternative distributions also address **non-constant variance** given that their variance is tied to the mean (can shrink as needed near boundaries)

- This removes the assumptions of residual normality and constant variance, but generalized models still assume **conditionally independent outcomes**…
  - For "**dependent**" outcomes, we need a **multivariate** model instead of univariate
  - Btw, we also use multivariate models to test differences in slopes across outcomes (i.e., does X → Y1 more than X → Y2?) or to test "mediation" (X → M → Y?)

From Univariate to Multivariate—Another use of ML

- **Multivariate models** (via **ML**) are needed in many common research designs:
  - You have **more than one outcome per person** created by multiple occasions and/or conditions (e.g., longitudinal or repeated measures designs)
  - When your **outcome is measured multiple times** for a pair or group (e.g., dyadic or family data)
  - More generally, multivariate models are used for an outcome that reflects **multiple dimensions of sampling, such as in multilevel** designs (e.g., students from different classes and/or schools)

- In multivariate GLMs (e.g., repeated measures ANOVA, MANOVA, MANCOVA), **OLS estimation quickly becomes useless**…
  - Does not allow any **missing** repeated measures (listwise-deletes entire person)
  - **Only two options for residual correlation** between outcomes (all same or all different)
  - Requires same **discrete occasions** in longitudinal data (i.e., balanced time)
  - No real RQs are likely answered by MANOVA, discriminant function, or canonical correlation (i.e., old school multivariate techniques we decided to fire from disuse in favor of modern Legos)

Multilevel Models (MLMs) for Clustered* Data

- **Clustering = Nesting = Grouping = Hierarchies*
  - "**Micro**" units are nested in one or more types of "**macro**" units
  - A "**level**" is a sampling dimension with variation remaining after prediction by fixed effects
  - More generally, **MLMs are multivariate models on "macro" units** used to quantify and predict distinct sources of outcome variance arising from each dimension of sampling

- The term "Multilevel Model" (MLM) has many synonyms:
  - **General(*ized*) Linear Mixed-Effects Models** (Fixed effects + Random effects = Mixed effects)
  - **Random Coefficients Models** (Random effects = latent variables)
  - **Hierarchical Linear Models** (HLM, but not the same as hierarchical regression)
  - Most MLM software is "univariate" → predict 1 conceptual outcome at a time
  - Multivariate MLMs can be estimated as "multilevel structural equation models" to predict 2+ conceptual outcomes at once (+ address missing predictors)

Multilevel Model (MLM) Special Cases

- Random Effects ANOVA or Repeated Measures ANOVA
  - Occasions are nested within persons after controlling for mean differences
- (Latent) Growth Curve Model (where "Latent" implies SEM software)
  - Btw, most longitudinal MLMs can be equivalently estimated as single-level SEMS
- Within-Person Fluctuation Model (e.g., for EMA or intensive longitudinal data)
  - See also "dynamic" SEM or multilevel SEM (even without measurement models!)
- Clustered/Nested Observations Model (e.g., for students within schools)
  - If followed over time in same group, is "clustered longitudinal model"
- Cross-Classified Models (e.g., teacher "value-added" models)
  - See also subjects crossed with items, raters crossed with targets
- Psychometric Models (e.g., factor analysis, item response theory, SEM)
  - Items modeled as nested in persons (or crossed, as in explanatory IRT)

Intro to MLMs for Clustered Data (EDF 9850)

- An "empty" two-level model for level-1 person $\boldsymbol{p}$ in level-2 cluster $\boldsymbol{c}$:

  Level-1: $\boldsymbol{y_{pc} = \beta_{0c} + e_{pc}}$

  Level-2: $\boldsymbol{\beta_{0c} = \gamma_{00} + U_{0c}}$

  $\boldsymbol{\gamma_{00}}$ = fixed intercept (mean of cluster means)
  $\boldsymbol{U_{0c}}$ = level-2 random intercept (with variance $\boldsymbol{\tau^2_{U_0}}$)
  $\boldsymbol{e_{pc}}$ = level-1 residual (with variance $\boldsymbol{\sigma^2_e}$)

- **Total** outcome variation is partitioned into **two uncorrelated sources**:
  - **Level-2 between**-cluster (BC) mean differences → random intercept $\boldsymbol{\tau^2_{U_0}}$
  - **Level-1 within**-cluster (WC) cluster differences → residual $\boldsymbol{\sigma^2_e}$
  - Dependency effect size via Intraclass Correlation: $\mathbf{ICC} = \boldsymbol{\tau^2_{U_0}}$ **/** $\boldsymbol{(\tau^2_{U_0} + \sigma^2_e)}$
    - ICC = proportion of total variance due to cluster mean differences (= within-cluster correlation)

- **Level-2** predictors explain **cluster mean** differences (reducing intercept variance $\boldsymbol{\tau^2_{U_0}}$)
- **Level-1** predictors explain **people** differences (reducing residual variance $\boldsymbol{\sigma^2_e}$)
- **Cross-level** interactions explain **cluster** differences in slopes of people predictors

That's a Wrap!

- General linear models are for predicting a single conditionally normally-distributed outcome (with constant variance) in an independent sample
  - **Not constant variance**? Likely need heterogeneity-corrected or Bootstrap standard errors
  - **Not normal residuals**? Likely need Generalized linear model family instead (or maybe quantile regression, especially to address outliers)
  - **Not independent sample** (so is dependent instead)? Likely need multivariate models (with multilevel models as a special case with much general utility)
  - **Not perfectly reliable measures** (or different measures across people)? Likely need latent variable measurement (psychometric) models through structural equation modeling

- But—good news—these options will require all the linear modeling and programming skills you have acquired this semester (and then build on them with additional Lego blocks)
  - **THANK YOU for all your efforts—I hope this wasn't *as* bad as you may have feared! ☺**