The Finale of PSQF 6243: Caveats and Next Steps

- Topics:
  - Summary of what we've covered as "The GLM"
  - Using what we've covered:
    - Review of steps in GLM analysis
    - Understanding GLM assumptions
  - What to do given untenable assumptions:
    - Repairs within GLM framework
    - When to eject to a new model (and what it could be)

Two Reasons Why You WERE Here

1. "This class fulfills a requirement" (and I just need to pass it).
   - I get it—it's ok if this is the *only* reason you were here, but I (still) hope to have convinced you otherwise!

2. "I want to learn more about data analysis using **quantitative methods**" (yes, me too)!
   - **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, syntax) to logic (matching data types, questions, 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… but first, let's review…

Reviewing the 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: *dependent variable*, *criterion*

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

Reviewing the Steps in a GLM Analysis

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

- **Quantitative variables** should be **centered** (by 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
  - 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** (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—the data can help you decide!

Reviewing the Steps in a GLM Analysis

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

- **Categorical predictors** (numbers are just **labels**) can only be included as-is if they have only 2 values (binary 0 or 1)
- Otherwise, they need to be represented by $C - 1$ new "dummy-coded" binary predictors for C categories *(*which can be done for you by using a "factor" variable, see unit 7)*

Indicator Coding* (best 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 (best 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**

- Btw, there is also "effect coding" (using ± 1), which I don't like because then 0 does not indicate any specific category

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 single slopes** using linear combinations of fixed intercept and slopes
     - e.g., other predictor category differences (such as B v D)
     - e.g., differences between fixed slopes (such as between piecewise slopes or predictor slopes on same scale)
     - e.g., conditional slopes for predictors also included in interaction terms at other moderator values besides 0
   - Simultaneous F-tests (DF>1) of **slopes lumped together**
     - e.g., for "omnibus" effects of categorical predictors
     - e.g., for testing changes to the model R^2 for a set of new slopes (avoiding hierarchical models in which order is ambiguous)

Reviewing the Steps in a GLM Analysis

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

   - Always ok: **per-slope partial d or r** → 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 (b/c 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 dummy-coded predictors to represent a categorical variable need to be lumped together first (b/c not independent)
     - Slopes of predictors involved in interaction terms (including quadratic terms) are conditional on moderators = 0

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—choose those that tell the honest story in answering your research questions, including:
  - **Analytic method**: modeling approach, software used, how predictors were centered or coded (i.e., who the reference is)
    - Equations are useful when done correctly—use the proper notation!
  - **What happened**: in both "stat-ese" and regular language (see my results sections in class examples and in homework assignments)
    - 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
  - Consider "**supplemental material**" for any results you don't have room for, as well as equations and syntax—can help you get cited!

Labels for What We Covered This Semester

Intro to **General Linear Models** (GLMs) as a one-stop shop *for predicting one conditionally normal outcome per person*

- **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, piecewise) 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!
  - See lecture 7 and Example 7 (posted from PSQF 6242) for interactions involving 2+ slopes (such as among predictors with 3+ categories or piecewise slopes)

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 effect **estimates**
    - Wrong estimates → wrong depiction of variable relations
    - Labeled "primary" by Darlington & Hayes (2017)
  - On fixed effect **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…

Is "Linearity" an Assumption?

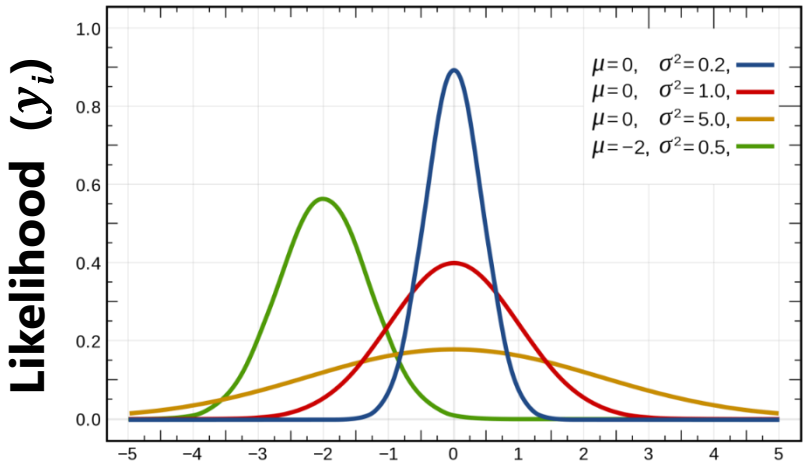
- In a word, **NO**! This is a misconception given that what are most commonly specified for quantitative predictors are linear slopes

- 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
  - "**Nonlinear in the parameters**" means a model that does not use the "constant*variable + constant*variable linear form
    - e.g., a truly exponential model: $y_i = \beta_0 + \beta_1[e^{\beta_2(x_i)}]$

- The GLM (and any model!) assumes the functional form of the predictor relations has been properly specified, 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)

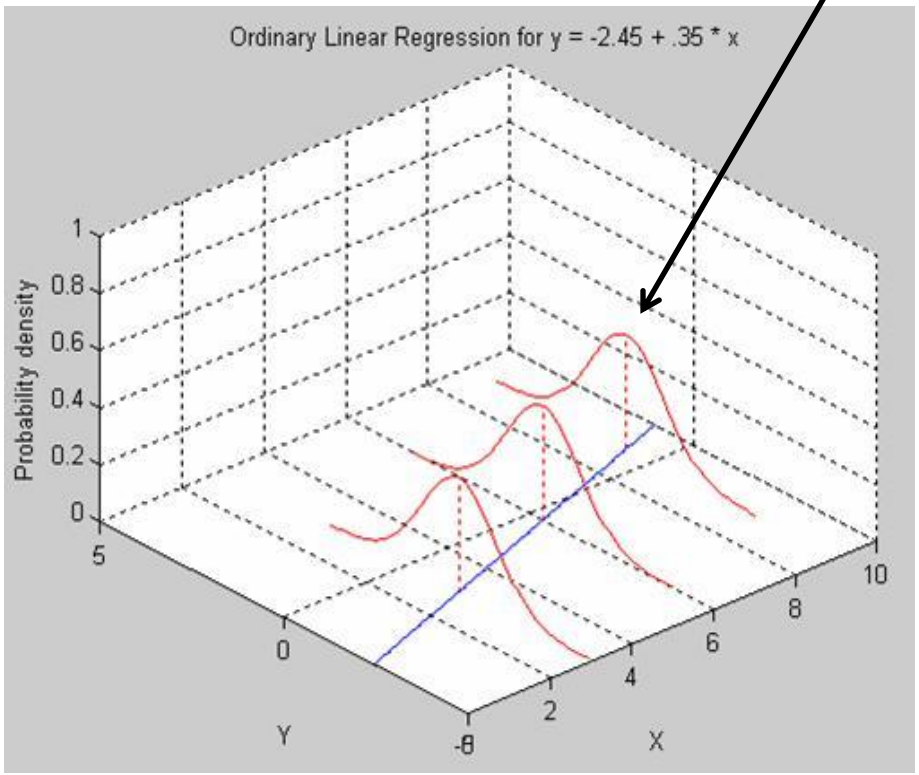


- Stand-alone textbook chapters on "data cleaning" and "data transformation" and "outlier analysis" are really problematic!
  - Because residuals are only possible in the context of a model!

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

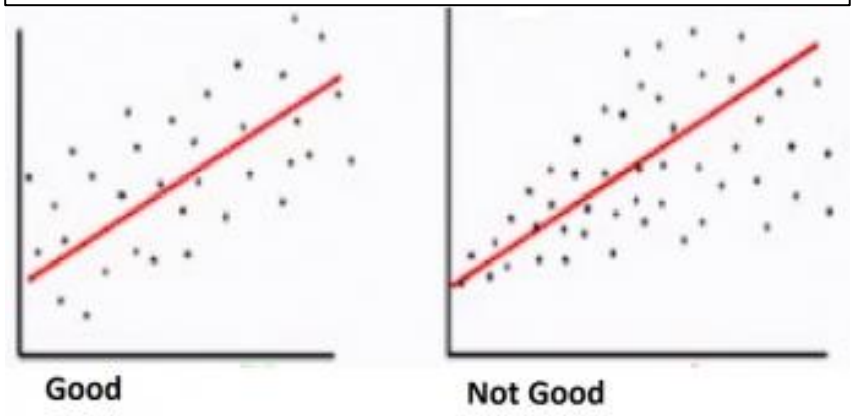
Normal Distribution → Constant Variance

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



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

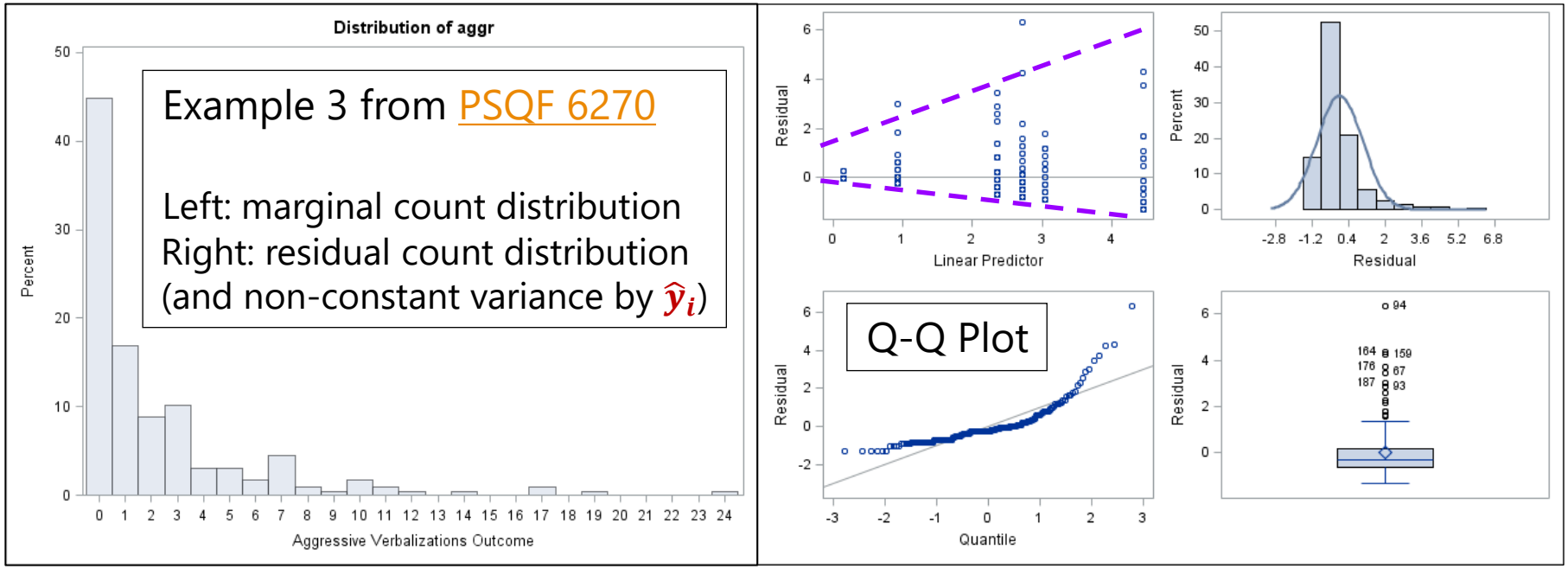
"Not good" → $\boldsymbol{\sigma_e^2}$ increases as the x_i predictor increases (→ fan shape)



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 one or more boundaries
  - e.g., proportion correct outcomes are bounded by both 0 and 1, so residual variance will shrink as $\hat{y}_i$ approaches these ends
  - e.g., count outcomes are bounded at 0, so residual variance usually increases with the predicted $\hat{y}_i$ (see example below)

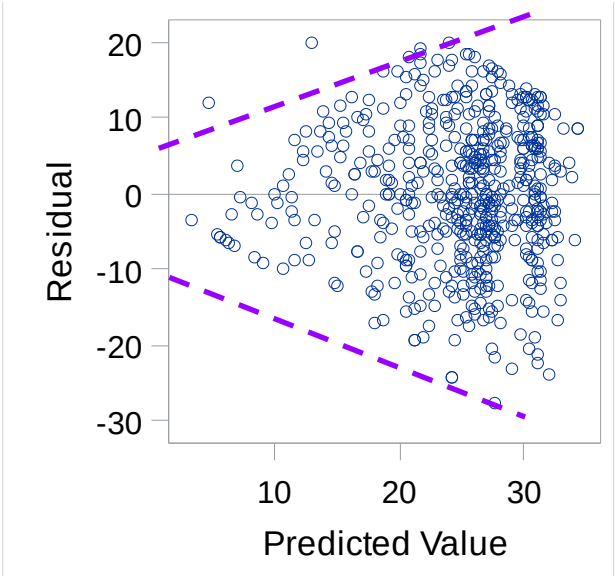


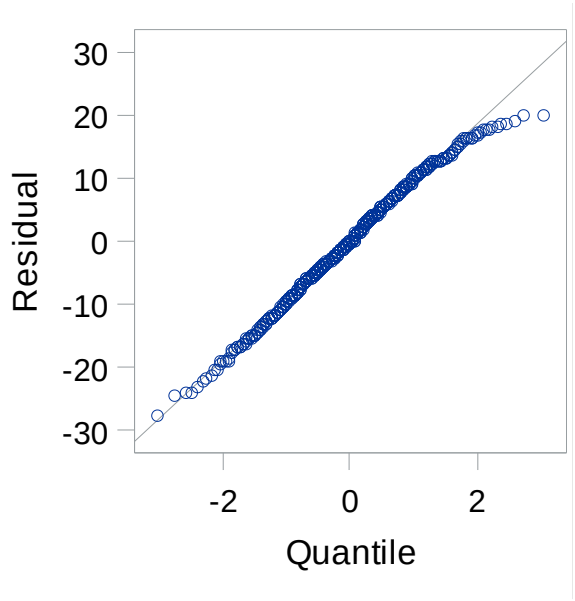
What to Do about Non-Constant Variance

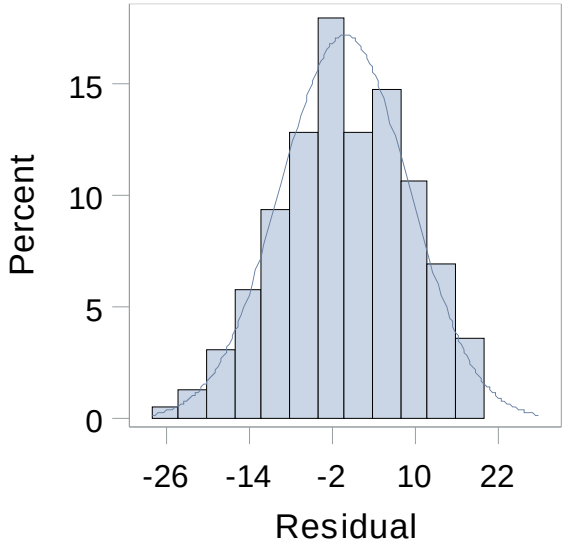
- Residual non-normality is not a big problem by itself
- In contrast, **non-constant residual variance** can result in incorrect standard errors (SEs) and thus **incorrect p-values**
- For outcomes that are "sufficiently continuous", the impact of non-constant variance can be addressed by **changing the way the standard errors (SEs) are computed**:
  - Request **heterogeneity-consistent** SEs
    - aka, "**sandwich**" estimators: predictors are the "bread"; weighted residuals are the "meat" (many types available, labeled HC0-HC4; I will show HC3 next)
  - Request jackknifed or bootstrapped SEs
    - **Jackknifing**: remove one person, re-estimate model N times, get empirical SD of estimates across resamples as new empirical SE
    - **Bootstrapping**: sample same N with replacement repeatedly, re-estimate model, get empirical SD of estimates across resamples as new empirical SE (more readily available)

Residuals for Last Model of Example 5

- Chapter 2 data: cognition outcome was simulated to have a normally-distribution residual e_i with constant variance
  - Plots show plausible normality, but non-constant variance *(because a sex*dementia group interaction is still missing relative to the correct population model, see Example 7!)*







Chapter 2 Data: Robust SEs (in SAS)

- Chapter 2 data: cognition outcome was simulated to have a normally-distribution residual with constant variance, so results don't change much when using robust SEs (last columns)
  - Biggest difference is for dementia slopes (which are mis-specified!)

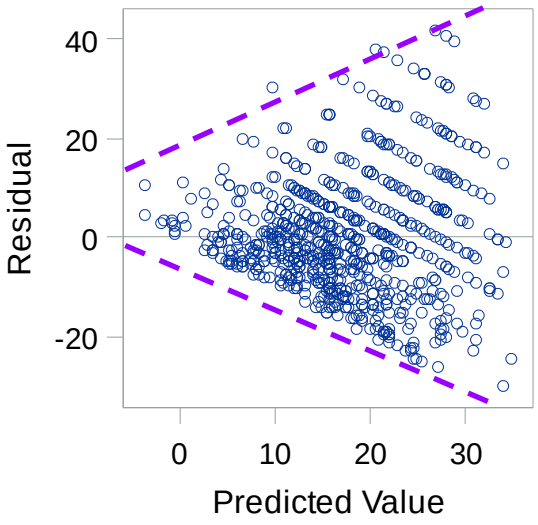
```
TITLE1 "SAS Chapter 2 Data: REG with Robust Standard Errors";
PROC REG DATA=work.Chapter2;
  MODEL cognition = age85 grip9 sexMW demNF demNC age85grip9
                    / ALPHA=.05 HCCMETHOD=3 WHITE;
RUN; QUIT; TITLE1;
```

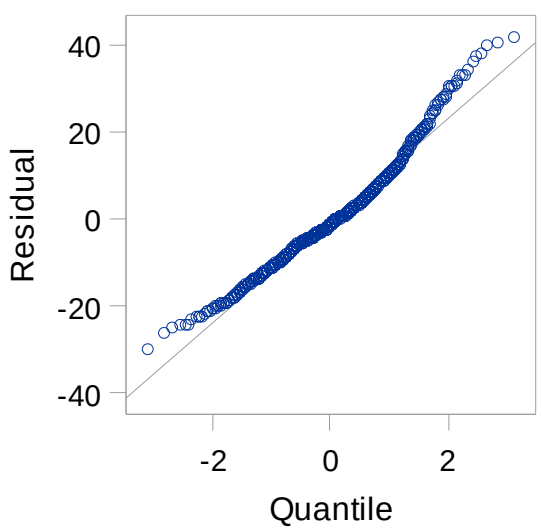
Parameter Estimates

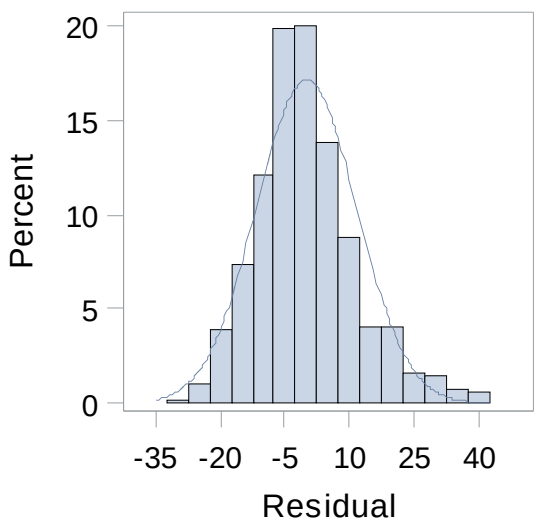
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Heteroscedasticity Consistent Standard Error	Heteroscedasticity Consistent t Value	Heteroscedasticity Consistent Pr > |t|
Intercept	Intercept	1	29.40780	**0.69491**	42.32	<.0001	**0.70729**	41.58	<.0001
age85	age85: Age in Years (0=85)	1	-0.33396	**0.12036**	-2.77	0.0057	**0.12009**	-2.78	0.0056
grip9	grip9: Grip Strength in Pounds (0=9)	1	0.61942	**0.14874**	4.16	<.0001	**0.15399**	4.02	<.0001
sexMW	sexMW: Sex (0=M, 1=W)	1	-3.45564	**0.88727**	-3.89	0.0001	**0.92128**	-3.75	0.0002
demNF	demNF: Dementia None=0 vs Future=1	1	-5.92254	**1.01363**	-5.84	<.0001	**1.02380**	-5.78	<.0001
demNC	demNC: Dementia None=0 vs Current=1	1	-16.30040	**1.51255**	-10.78	<.0001	**1.16189**	-14.03	<.0001
age85grip9	age85*grip9: Age by Grip Interaction	1	0.12302	**0.04054**	3.03	0.0025	**0.04236**	2.90	0.0038

Residuals for Last Model in Example 4a

- GSS real data: income was binned (so it's not really continuous) and has a natural lower boundary at 0
  - Residual plots show both (some) non-normality and non-constant variance across predicted outcomes







GSS Data: New SEs Differ! (in STATA)

```
display "STATA GSS Data: REGRESS with Robust Standard Errors"
regress income c.LvM c.LvU c.age18 c.age18#c.age18 ///
        c.lessHS c.gradHS c.overHS, level(95) vce(hc3)
```

income	Coef.	Std. Err.	t	P>|t|
LvM	6.060105	.9470067	6.40	0.000
LvU	7.208538	2.697879	2.67	0.008
age18	1.06998	.1230046	8.70	0.000
c.age18#c.age18	-.0175062	.0022749	-7.70	0.000
lessHS	.2589179	.5612016	0.46	0.645
gradHS	3.157139	1.757267	1.80	0.073
overHS	1.528179	.2080423	7.35	0.000
_cons	-3.686546	2.004615	-1.84	0.066

income	Robust HC3 Coef.	Std. Err.	t	P>|t|
LvM	6.060105	.9917771	6.11	0.000
LvU	7.208538	4.180355	1.72	0.085
age18	1.06998	.1032716	10.36	0.000
c.age18#c.age18	-.0175062	.0021002	-8.34	0.000
lessHS	.2589179	.3759458	0.69	0.491
gradHS	3.157139	1.192412	2.65	0.008
overHS	1.528179	.222025	6.88	0.000
_cons	-3.686546	1.37495	-2.68	0.008

Top left: original results
Top right: with robust SEs
Bottom: with bootstrapped SEs

```
vce(bootstrap, reps(500))
```

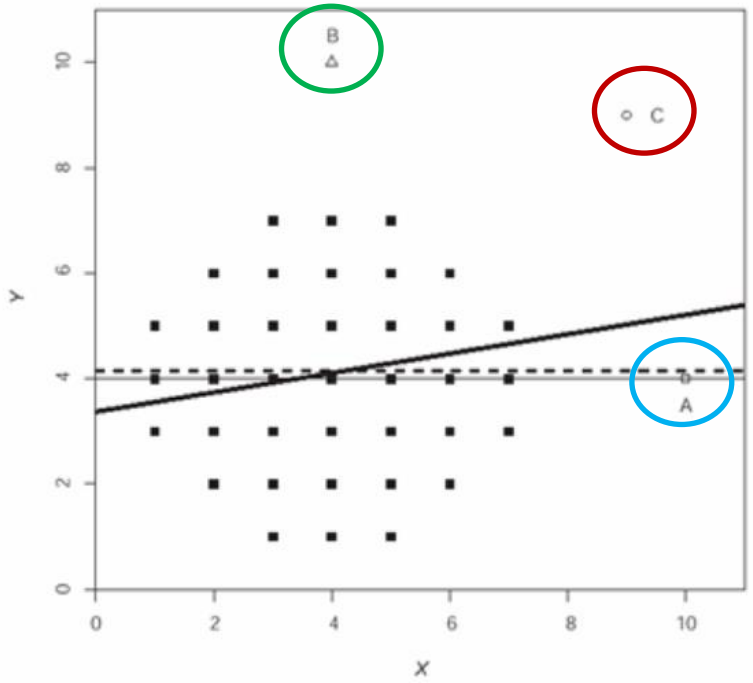
Btw, an R example is given here

income	Observed Coef.	Bootstrap Std. Err.	z	P>|z|
LvM	6.060105	1.027345	5.90	0.000
LvU	7.208538	4.344758	1.66	0.097
age18	1.06998	.10824	9.89	0.000
c.age18#c.age18	-.0175062	.0021987	-7.96	0.000
lessHS	.2589179	.3680929	0.70	0.482
gradHS	3.157139	1.21188	2.61	0.009
overHS	1.528179	.2130186	7.17	0.000
_cons	-3.686546	1.380646	-2.67	0.008

Problematic Participants

- Skewed residuals can also be due to extreme values ("outliers")

- "**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 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…

Image 16.1 from Darlington & Hayes (2017)

When to Eject to a New Model Family

- For some outcome variable formats, the assumptions of normally-distributed residuals with constant variance will **never** be plausible
  - Then you need a **general*ized* linear model** instead of a general linear model (where "ized" → not normal)
  - These models swap a normal residual distribution with a more plausible distribution and build in a link function transformation to keep predicted outcomes within their possible bounds (which is necessary when predicting categorical, count, or skewed positive continuous outcomes)

- I created PSQF 6270: Generalized Linear Models as a follow-up to PSQF 6242/6243 for this reason!
  - Btw, it also covers **multivariate models** for predicting multiple outcomes at once, as well as **path analysis** for testing mediation
  - Btw, it also covers "**quantile regression**" in which you can predict the median (or any percentile) instead of the mean to avoid bias in results due to cases with strong potential influence

When to Eject to a New Model Family

- The most important assumption of the GLM 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
  - First "i" in assumption abbreviation **"i.i.d"**: residuals are **i**ndependently and **i**dentically **d**istributed (→ unrelated with constant variance)
  - 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)

- **Ignoring correlated residuals can lead to WAY-wrong fixed effect estimates, SEs, and p-values for several reasons!**
  - **Effective sample size** is lower than actual N, so SEs need to adjust for redundancy created by to residual correlation
  - Each lower-level predictor is **really two predictors in one** (with distinct effects across each dimension of sampling)

Ejecting to a New Model Family

- Research designs with random variation across multiple dimensions of sampling simultaneously need linear models that can allow residuals from the same sampling unit to be correlated (and have non-constant variance)

- For some cases, you can introduce correlated residuals directly (as in multivariate models, which are covered in PSQF 6270)
  - e.g., pre-test and post-test for same people for a conditionally normal outcome

- Otherwise, you need "**multilevel**" models (MLM): here is a recent tutorial
  - aka, "**mixed-effects**" models (where "mixed" means inclusion of fixed and random effects) or "**hierarchical linear models**" (HLM)
  - Introduced in PSQF 6246 Design of Experiments
  - **Repeated measures** designs covered in new PSQF 6271: Longitudinal Multilevel Models
  - **Clustered/nested** designs covered in new PSQF 6272: Clustered Multilevel Models

- Require **likelihood estimation** instead of ordinary least squares to address missing data and/or unbalanced designs (different possible outcomes per sampling unit)
  - Which is why I didn't cover "traditional" dependent samples t-tests or repeated measures analysis of variance in this class…
  - Likelihood estimation is introduced in PSQF 6270 (so take that before 6271 or 6272)

What about Measurement Error?

- Any linear model using observed variables (GLM, generalized, or multilevel) assumes they are measured with perfect reliability
  - Highly unlikely in social sciences examining "squishy" constructs
  - We frequently compute reliability coefficients (e.g., alpha, which is useless, btw), but then we don't address reliability in the linear models!

- But measurement error can reduce the size of the variable relations captured and otherwise introduce bias in slopes
  - e.g., covariates will not be adequately "controlled for" if they are not measured well to begin with!

- Psychometric models use **latent variables** to more reliably measure an unobserved construct than can any single observed variable
  - e.g., using multiple item responses to measure a latent construct
  - Learn about these in PSQF 6262: Item Response Theory (IRT) and PSQF 6249: Factor Analysis and Structural Equation Models (SEM) as well as various special topics courses (listed as PSQF 7375 or 7476)

Wrapping Up

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

- But good news—these options will require all the linear modeling and programming skills you have acquired this semester (and build on them)
  - **THANK YOU for all your efforts—I hope to see you in class again!**