Categorical Data Analysis

Table of contents

1 Goals

1.1 Goals

1.1.1 Goals of this section

- Review linear regression
  - **Assumptions** and how categorical variables violate them
- Introduce **generalized linear model (GLiM)** framework

- Specific models in the GLiM family
  - Logistic regression
  - Ordinal and multinomial logistic regression
  - Poisson regression (inc negative binomial regression)
  - A few others, time permitting

1.1.2 Goals of this lecture

- What does "categorical" mean?
  - Levels of measurement of variables
- Linear regression
  - Assumptions
  - Violation of assumptions
- Generalized linear model (GLiM) framework

2 Levels of measurement

2.1 Levels of measurement

2.1.1 Continuous vs categorical?

- We talk about "continuous" variables or "categorical" variables
  - Sometimes the distinctions between them are easy to see
  - But often they are not
- We are going to talk about **levels of measurement** for variables
  - A more fine-grained, *nuanced* discussion of types of variables
- Focus on **why it matters**

2.1.2 Levels of measurement

- Attributed to Stevens (1946)
- Four **ordered** levels of measurement
  - Nominal
  - Ordinal
  - Interval
  - Ratio

2.1.3 Nominal variables

- **Categories** with **no intrinsic ordering**
  - Nominal = "name"
- Examples
  - Department: Psychology, Epidemiology, Statistics, Business
  - Religion: Christian, Jewish, Muslim
  - Ice cream flavor: vanilla, chocolate, strawberry

2.1.4 Ordinal variables

- **Categories** with **some intrinsic ordering**
  - Ordinal = "ordered"
  - Differences between categories are **not meaningful**
- Examples
  - Dose of treatment: low, medium, high
  - Rankings: 1st, 2nd, 3rd, 4th
  - Education: high school, some college, college grad, graduate
  - Likert scales: agree, neutral, disagree

2.1.5 Interval variables

- **Quantitative variables** with **no meaningful 0 point**
  - ("Meaningful 0": value of 0 = nothing)
  - **Differences** between values are meaningful but **ratios** are not!
- Example: Temperature in Fahrenheit or Celsius
  - **Difference** from 100F to 90F = **difference** from 90F to 80F
  - But 100F is **not twice** 50F (because 0F is arbitrary)
- Most "continuous" variables you deal with are **interval**
  - Most statistical procedures assume interval-level measurement

2.1.6 Ratio variables

- **Quantitative variables** with **meaningful 0 point**
  - ("Meaningful 0": value of 0 = nothing)
  - **Differences** between values are meaningful and so are **ratios**!
- Example: Temperature in Kelvin
  - **Difference** from 100K to 90K = difference from 90K to 80K
  - 100K is **twice as hot** as 50K (0K is *zero* molecular movement)
- Few variables in the behavioral sciences are ratio-level
  - Age, weight

2.1.7 Summary of levels of measurement

1. Nominal: unordered categories
2. Ordinal: ordered categories
3. Interval: quantitative with no meaningful 0 point
4. Ratio: quantitative with meaningful 0 point

2.1.8 Stevens (1946)

The *levels of measurement* determines what **mathematical** (and **statistical**) operations you can perform

Mathematical operation	Nominal	Ordinal	Interval	Ratio
equal, not equal	✓	✓	✓	✓
greater or less than		✓	✓	✓
add, subtract			✓	✓
multiply, divide				✓
central tendency	mode	median	mean	mean

2.1.9 Categorical outcomes

- Most "categorical" variables are nominal or ordinal
  - **Binary** variables (e.g., yes / no)
  - **Ordered categories** (e.g., Likert items)
  - **Unordered categories** (e.g., race / ethnicity)

- **Counts** are considered categorical but are *ratio*

- ANOVA and regression models focus on **means**
  - We can't calculate means for nominal or ordinal variables
  - What can we do?

3 General linear model (GLM)

3.1 General linear model (GLM)

3.1.1 General linear model (GLM)

- The general linear model (GLM) is
  - a "general" statistical model
  - to predict a *single continuous, conditionally normally distributed outcome variable*
  - from *one or more continuous or categorical predictors*

3.1.2 ANOVA and linear regression are GLM

- Analysis of variance (ANOVA) and linear regression (OLS regression) are both special cases of the general linear model
- ANOVA
  - 1 continuous outcome
  - 1 or more *categorical* predictors
- Linear regression
  - 1 continuous outcome
  - 1 or more *continuous or categorical* predictors

3.1.3 Linear regression

Two *equivalent* ways to present the linear regression equation

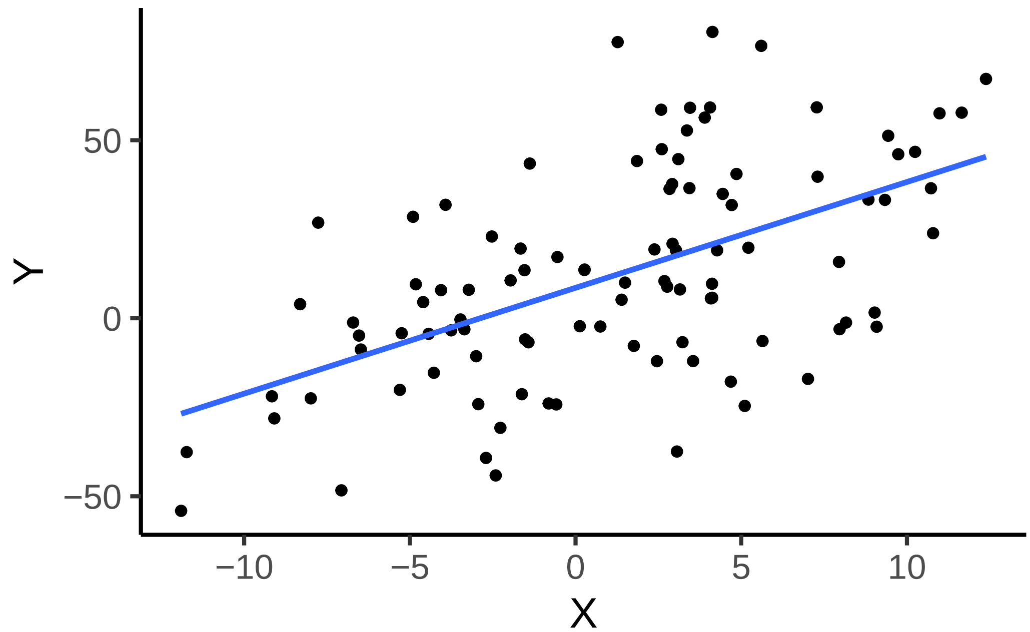
1. *Predicted score* is a fxn of coefficients and predictors, **no error term**

$$\widehat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi}$$

2. *Observed score* is a fxn of coefficients, predictors, **and error term**

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi} + e_i$$

3.1.4 Linear regression: $\hat{Y}$ = 8.55 + 2.97 X



3.2 Assumptions

3.2.1 Assumptions of GLM

- There are **three** major assumptions of GLM that are required to make *valid statistical inferences*
  - These assumptions are about the **residuals** of the model

1. Independence
2. Constant variance (homoskedasticity)
3. Conditional normality

3.2.2 Residuals

- Each subject has
  - **Observed** outcome value: Y_i

  - **Predicted** outcome value: $\hat{Y}_i$
  - **Residual** value: $e_i = Y_i - \hat{Y}_i$

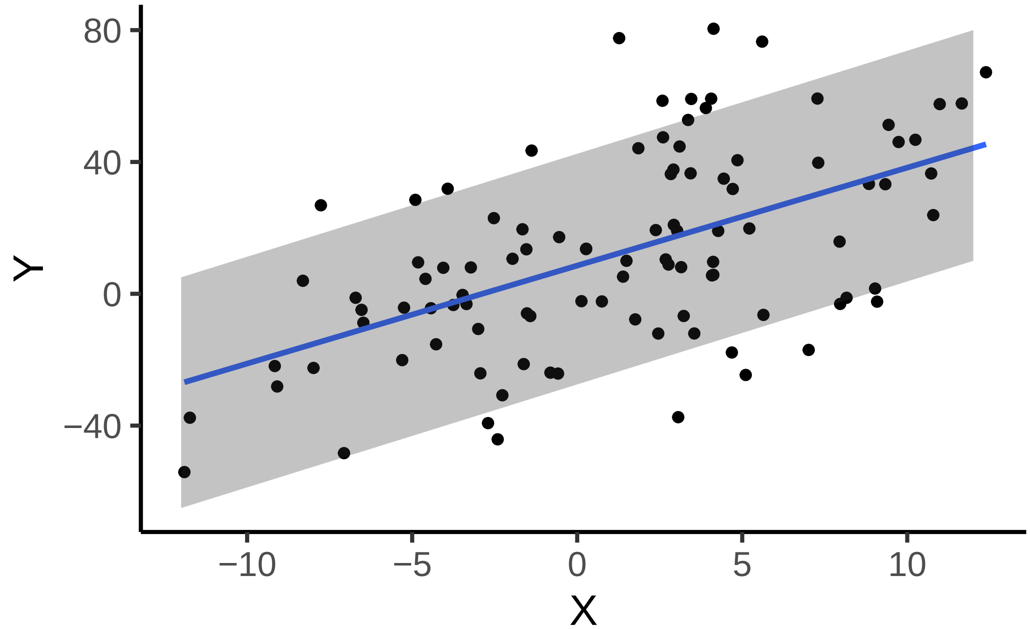
3.2.3 Independence

- **Independence of observations** means that subject i's values do not depend on subject j's values
  - Independent observations will be uncorrelated
  - But lack of correlation doesn't mean they're independent
- **Non-independence** occurs because of *clustering of observations in groups* (e.g., families, classrooms) or *repeated observations on the same person over time*
  - Not specific to categorical outcomes, but can always happen

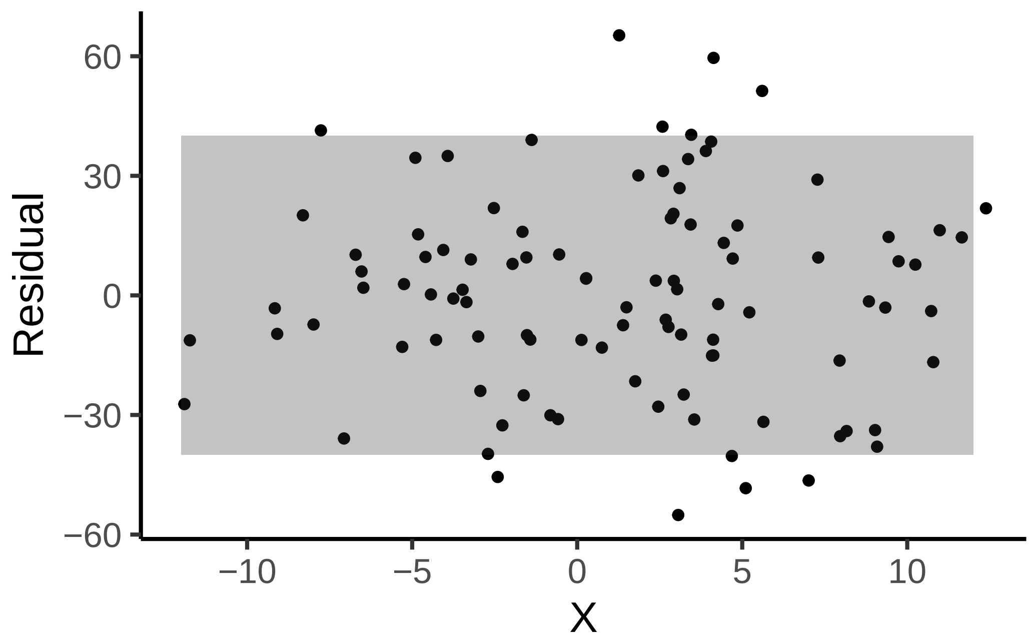
3.2.4 Constant variance

- **Homoskedasticity**: The variance of the residuals is **constant**, regardless of the value of the predictor(s)
  - *Heteroscedasticity* is the opposite (non-constant variance)
- **Any variable** can display heteroskedasticity
  - Categorical variables **typically** display heteroskedasticity
  - Binary variables (0,1) show increasing then decreasing variance
  - Count variables often show increasing variance

3.2.5 Constant variance



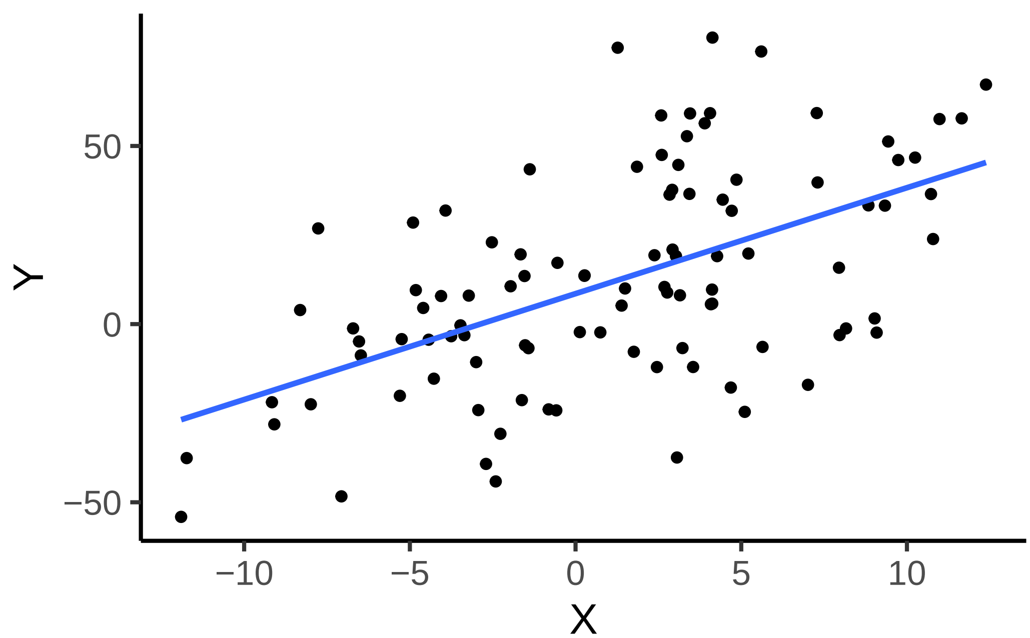
3.2.6 Constant variance



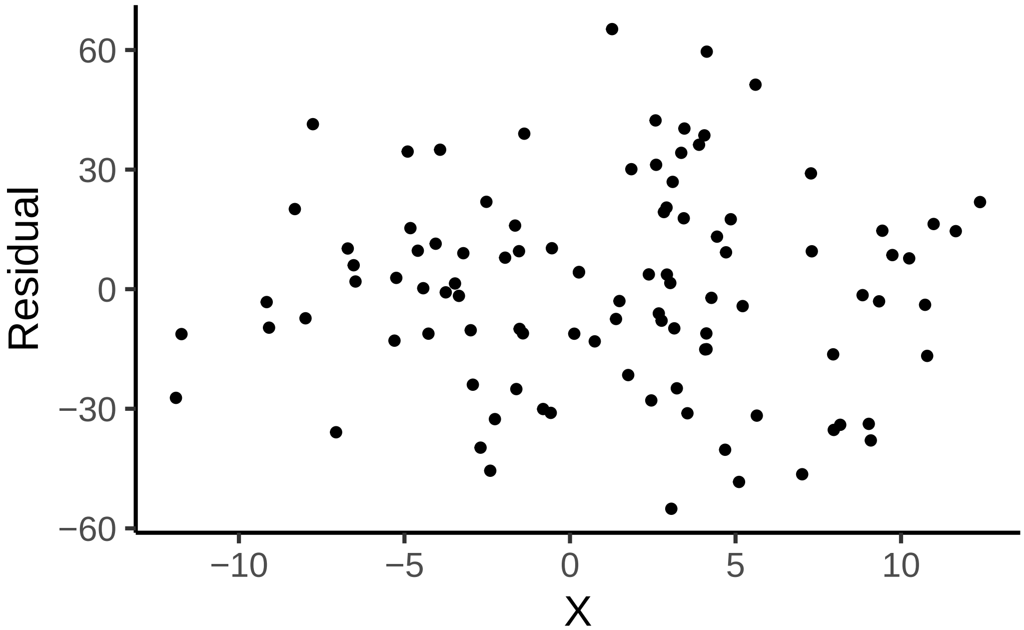
3.2.7 Conditional normality

- Residuals are normally distributed **at each value of the predictor(s)**
  - *Distribution of outcome variable* needn't be normal
  - *Overall distribution of residuals* needn't be normal
  - **Though one or both will often be true**
- Categorical outcomes often result in non-normal residuals
  - Often discrete and bounded
  - The normal distribution is **continuous** and **unbounded**

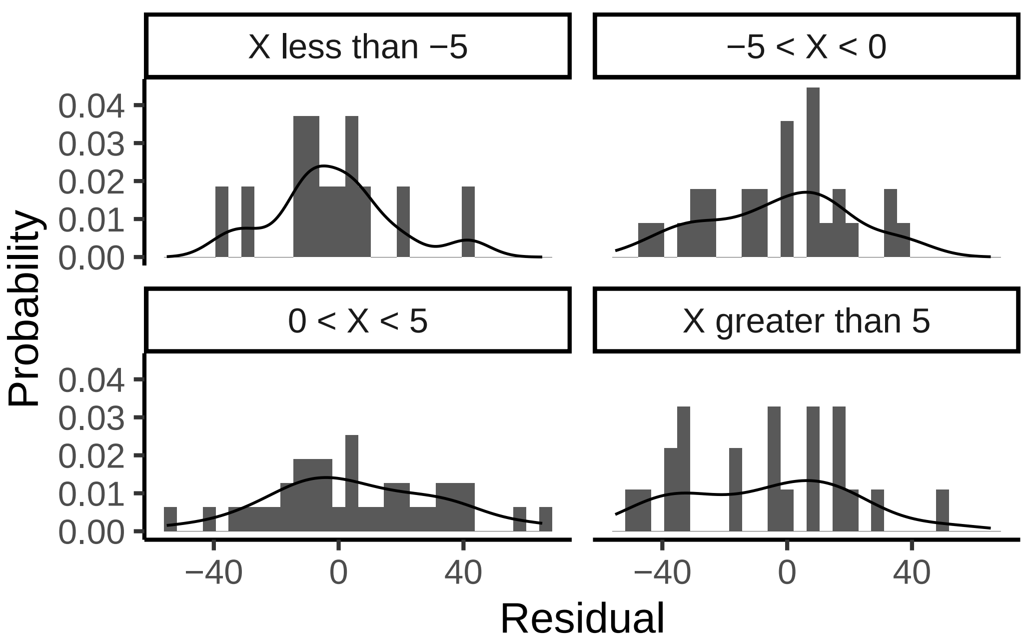
3.2.8 Conditional normality



3.2.9 Conditional normality



3.2.10 Conditional normality: Rough approximation

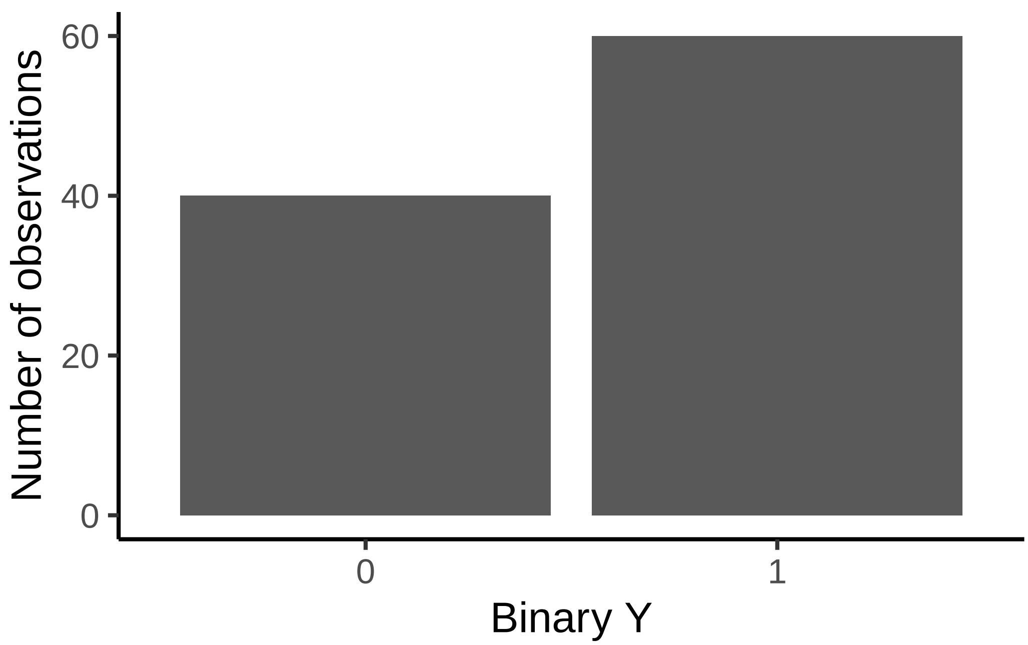


3.3 Violations of assumptions

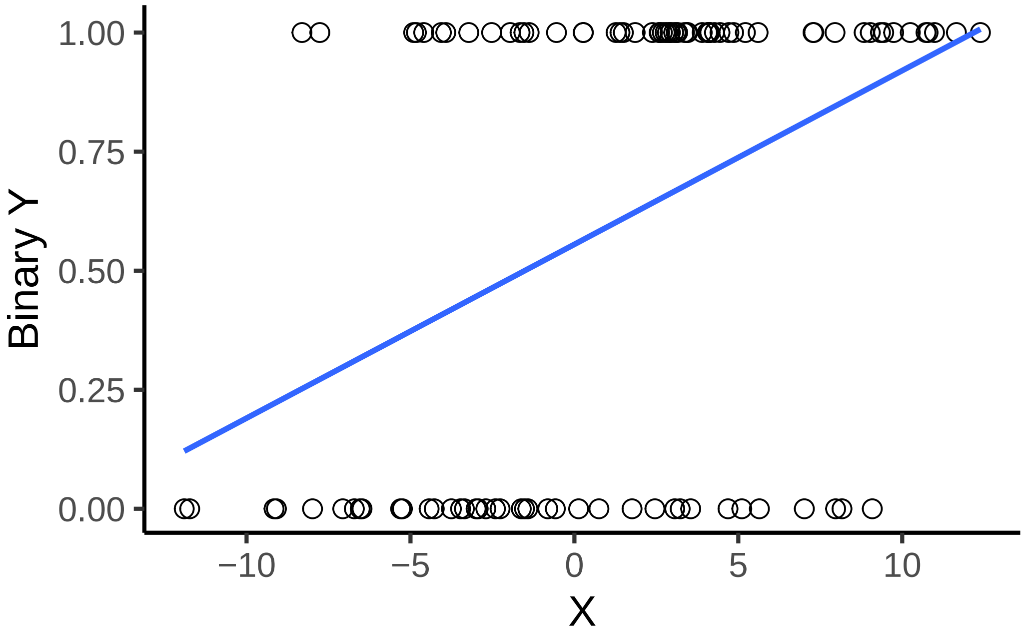
3.3.1 Violations: Non-normality of residuals

- Most statistical tests (such as t-tests of regression coefficients) are **parametric** tests that assume **normal distributions**
  - Non-normality of residuals means that these tests are not appropriate and will be **biased**
- I'm not referring to slight deviations from normality here
  - There is NO WAY to make a binary variable "approximately normal"

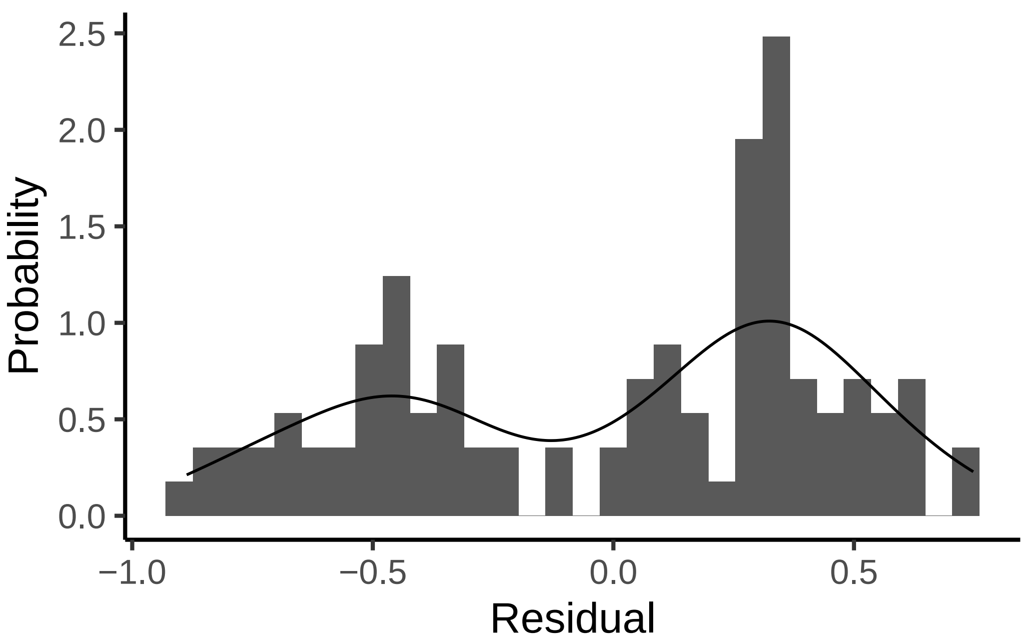
3.3.2 Violation of normality: Figures



3.3.3 Violation of normality: Figures



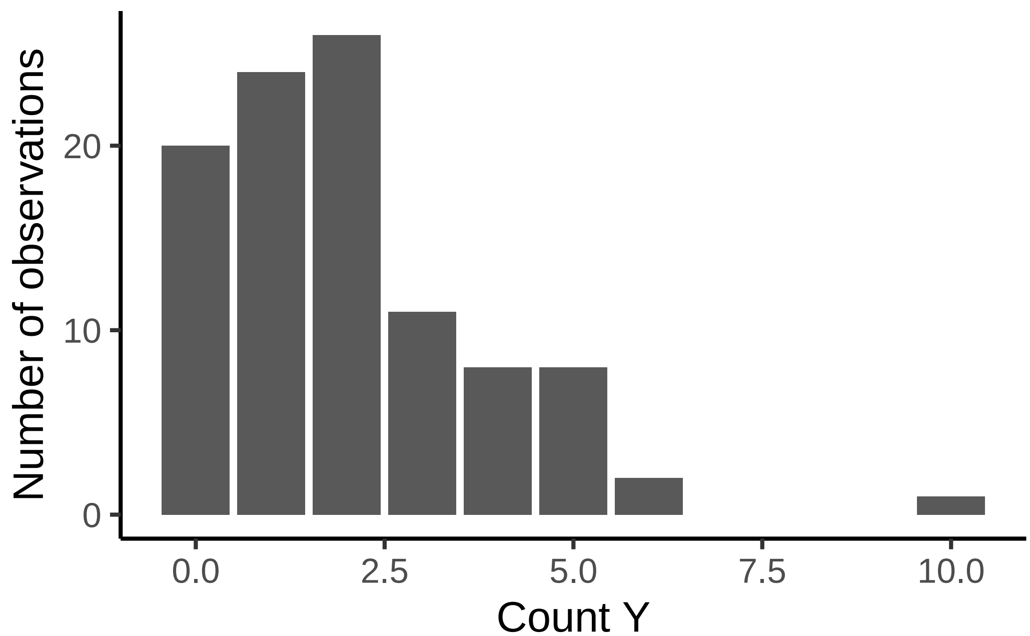
3.3.4 Violation of normality: Figures



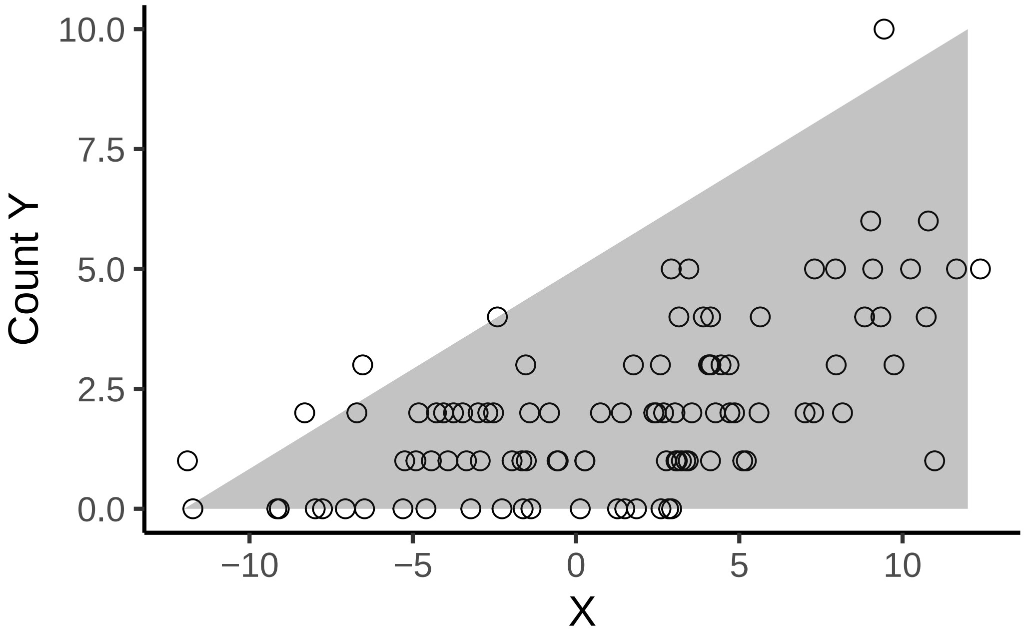
3.3.5 Violations: Heteroskedasticity

- Heteroskedasticity leads to **bias in standard errors**
  - Standard error may be *too high* or *too low*
- The t-test of a regression coefficient: $t = b/se_b$
  - where se_b is a function of the **constant** standard deviation of the residuals, σ
  - If the residuals have **non-constant variance**, there is not a single value of σ to use in calculating se_b

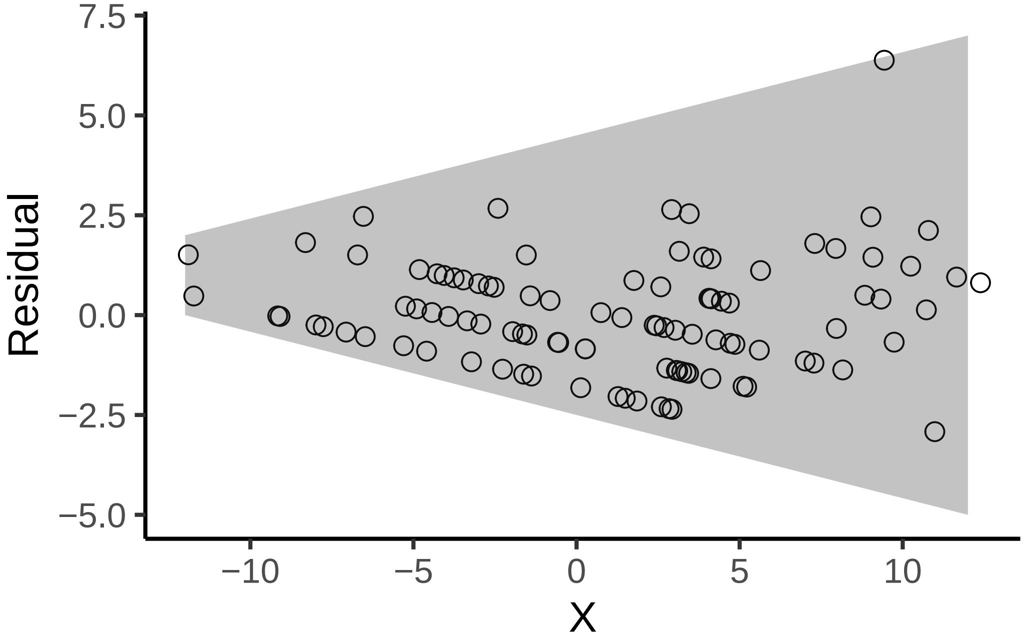
3.3.6 Violation of homoskedasticity: Figure



3.3.7 Violation of homoskedasticity: Figure



3.3.8 Violation of homoskedasticity: Figure

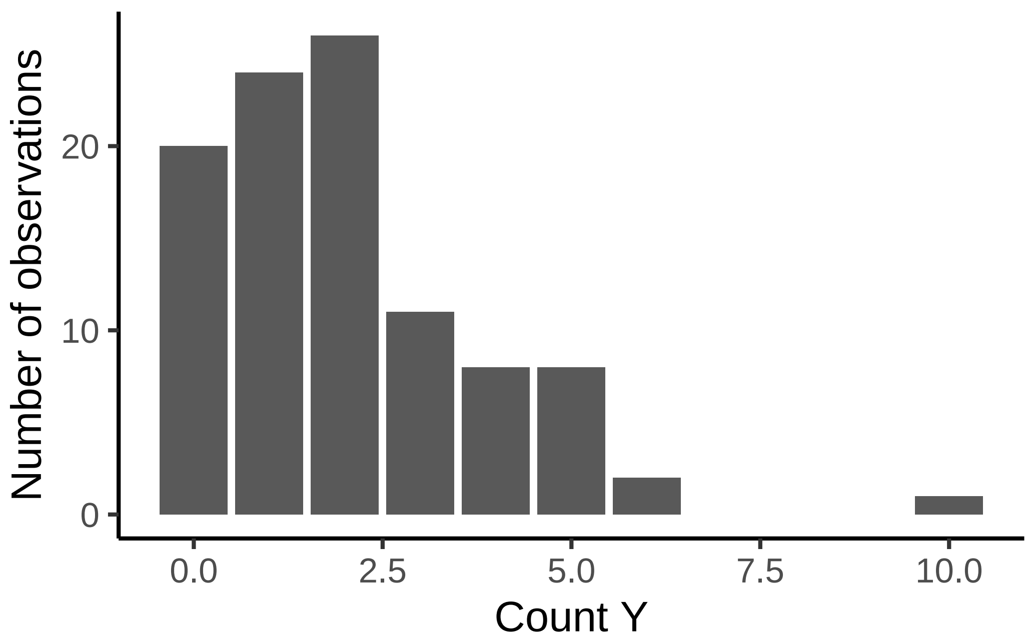


3.4 What NOT to do

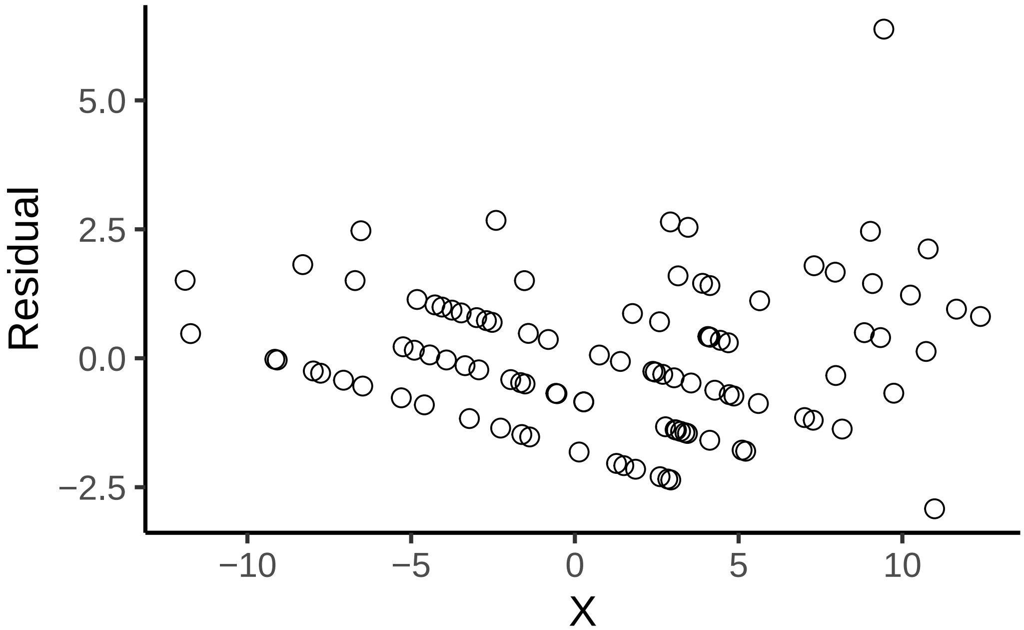
3.4.1 What NOT to do

- Historically, people have either **ignored** these violations or have used **transformations** of the outcome variable
  - e.g., natural log of a count, square root of a proportion
- **Problem**: *Transformations don't actually do what we think they do*
  - *May* slightly normalize the univariate distribution
  - But don't fix heteroskedasticity or conditional non-normality of residuals

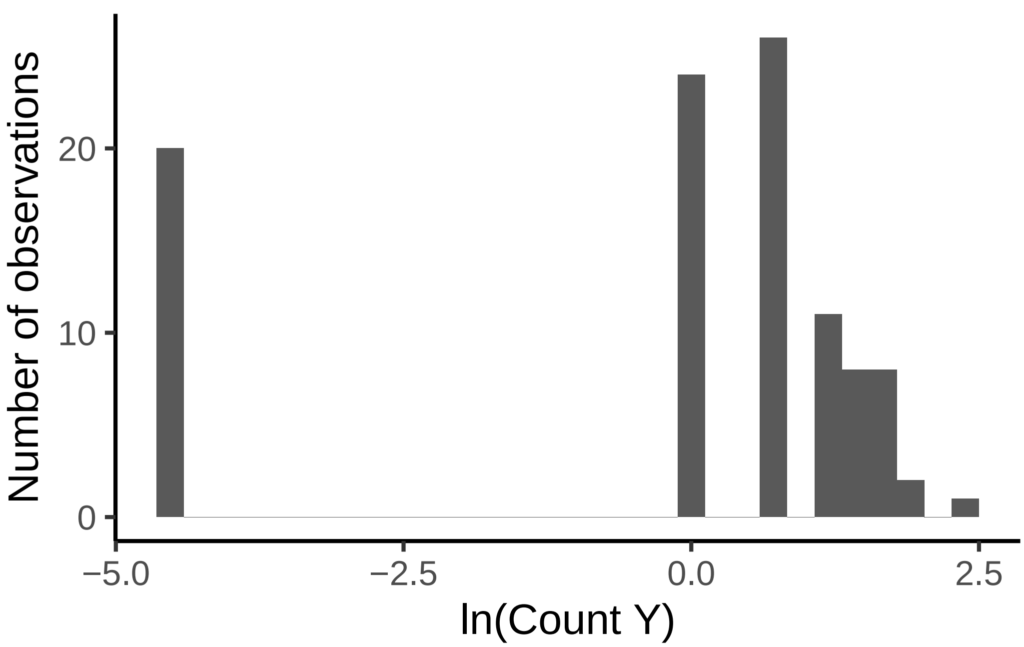
3.4.2 Count outcome



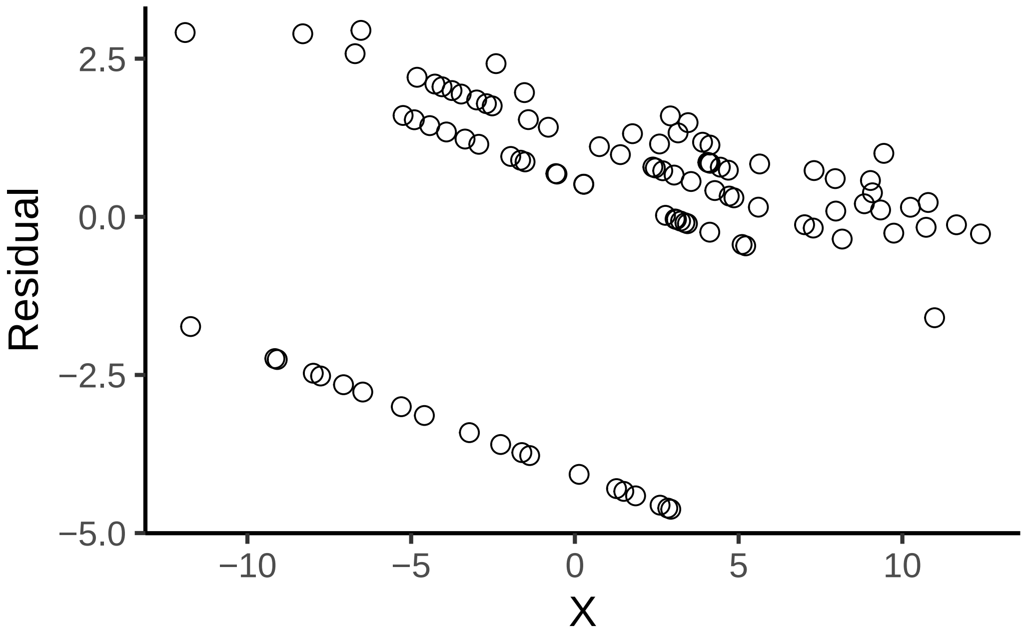
3.4.3 Residuals for count outcome in linear regression



3.4.4 Transform count: ln(count)



3.4.5 Residuals with ln(count)



4 Generalized linear model (GLiM)

4.1 Extension of GLM

4.1.1 GLiM is a "generalized" version of GLM

- Linear regression (GLM)
  - 1 continuous and conditionally normally distributed outcome
  - 1 or more continuous or categorical predictors
- Generalized linear model (GLiM)
  - 1 outcome that **may or may not** be continuous or conditionally normally distributed
  - 1 or more continuous or categorical predictors

4.1.2 GLiM family of models

- The generalized linear model (GLiM) is not just one model
  - It is a **family** of regression models

- Choose features (i.e., residual distribution) to match the characteristics of your outcome variable

4.1.3 GLiM framework

- All GLiMs have a similar underlying **framework**
  - *Random component*: distribution of the residuals
  - *Systematic component*: linear combination of predictors and regression coefficients
  - *Link function*: relates random and systematic components

4.1.4 Random component

- Distribution of residuals
  - Typically same as (conditional) distribution of the outcome
- GLiMs can use any distribution in the **exponential family**
  - Normal, exponential, binomial, multinomial, Poisson
    * All have $e^{something}$ in their probability distribution
    * Continuous outcome: (conditional) normal distribution
    * Binary outcome: (conditional) binomial distribution
    * Count outcome: (conditional) Poisson distribution

4.1.5 Systematic component

- In GLM, we talk about $\hat{Y}$, the expected or predicted value of Y
  - In GLiM, we will talk about η (eta), which is **some function of** $\hat{Y}$
  - (More on the "some function of" in a minute…)
- Specifically, we say that η is a function of the predictors (Xs) and regression coefficients (bs)
  - Also called the "linear predictor"
  - Systematic component: $\eta = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p$

4.1.6 Link function

- The link function relates $\hat{Y}$ (expected value of Y) to η
  - What needs to happen to get a straight line (systematic)
- Depends on the **outcome type** and **random component**
  - **You generally won't have any intuition about this**
  - *Canonical* links: Most commonly used, easiest to estimate
    - Identity link function: $\hat{Y} = \eta$
    - Logit link function: $logit(\hat{Y}) = ln(\frac{\hat{Y}}{1-\hat{Y}}) = \eta$
    - Natural log link function: $ln(\hat{Y}) = \eta$

4.1.7 GLiM example: Putting it together

- Continuous and normally distributed Y predicted by X
  - Systematic component: $\eta = b_0 + b_1 X$
  - Random component: Normal distribution
  - Link function: Identity ($\hat{Y} = \eta$)
  - Put them together:
    - $\hat{Y} = \eta = b_0 + b_1 X$
    - where the residuals $\sim N(0, \sigma^2)$

4.1.8 GLiM parts

- Even with this example, the three parts are probably a little *abstract* right now
- Next week, we'll talk about the *specific example* of **logistic regression**
  - That should make it more concrete
  - We'll also start talking more about **distributions** which should help with this idea of "picking a residual distribution"

4.1.9 Transformation of the predicted value

- I *just* told you **not to transform the outcome**, so ???
  - Notice that the link function uses $\hat{Y}$, not Y
  - Don't: **Transform** then *predict*
  - Do (using GLiM): **Predict** then *transform*

- For a linear transformation (add, subtract, multiply by a constant, identity), order doesn't matter
    - For a **non-linear** transformation (ln, logit, etc.), **order matters**

4.2 Similarities and differences

4.2.1 The same...

- t-test is a special case of ANOVA
- ANOVA is a special case of regression
- Linear regression and ANOVA are special cases of GLiM
    - GLiM with identity link and normally distributed residuals
- For a normally distributed outcome, you have a choice of using a *regression procedure* or using a *GLiM procedure*

4.2.2 ... with some differences

~	Linear regression	GLiM
Estimation	Ordinary least squares (OLS)	Maximum likelihood (ML)
Missing data	Listwise deletion	Maximum likelihood (ML)
Tests	t-tests	z or χ^2-tests*
Overall	R^2	Pseudo-R^2

* For a normal outcome, R gives t-tests in GLiM procedure

4.2.3 Linear regression procedure

```
model1_lr <- lm(y ~ x, data1)
summary(model1_lr)
```

```

Call:
lm(formula = y ~ x, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-55.087 -15.069  -0.278  15.475  65.242 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5513     2.5881   3.304  0.00133 **
x             2.9727     0.4557   6.523 3.03e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.27 on 98 degrees of freedom
Multiple R-squared:  0.3028,	Adjusted R-squared:  0.2957 
F-statistic: 42.56 on 1 and 98 DF,  p-value: 3.025e-09
```

4.2.4 GLiM procedure

```r
model1_glim <- glm(y ~ x, data1, family = gaussian(link = "identity"))
summary(model1_glim)
```

```

Call:
glm(formula = y ~ x, family = gaussian(link = "identity"), data = data1)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-55.087  -15.069   -0.278   15.475   65.242  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5513     2.5881   3.304  0.00133 **
x             2.9727     0.4557   6.523 3.03e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 638.64)

    Null deviance: 89764  on 99  degrees of freedom
Residual deviance: 62587  on 98  degrees of freedom
AIC: 933.7

Number of Fisher Scoring iterations: 2
```

5 Summary

5.1 Summary

5.1.1 Summary of this week

- What does “categorical” mean?
  - Levels of measurement of variables
- Linear regression
  - Assumptions
  - Violation of assumptions
- Generalized linear model (GLiM) framework

5.1.2 Next few weeks

- GLiMs that are used in psychology
  - Binary outcomes: Logistic (and probit) regression
  - Ordered categories (3+): Ordinal logistic regression
  - Unordered categories (3+): Multinomial logistic regression
  - Count outcomes: Poisson regression, overdispersed Poisson regression, negative binomial regression, excess zeroes versions of these models