Week 8: General(ized) linear models w/ different variable types

ANTH 674: Research Design & Analysis in Anthropology

Professor Andrew Du

Andrew.Du2@colostate.edu

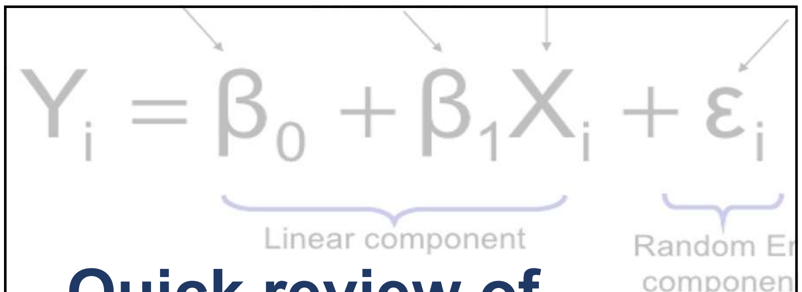
1

Lecture outline

1. Quick review of general linear models
2. Different types of GLMs (& their nonparametric counterparts)
   1. t-test
   2. ANOVA
   3. ANCOVA
   4. Logistic regression*
   5. Multinomial logistic regression*
   6. Chi-squared test*

*Technically, these are generalized linear models (non-normal errors)

2

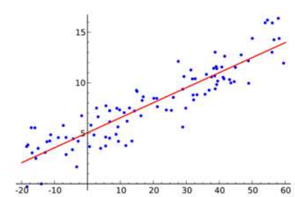


Quick review of general linear models

3

What are general linear models?

- Models continuous DV as a *linear/additive* function of one or more IVs
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_n X_n + \varepsilon$
- IVs can be continuous or categorical (so far, we have just covered continuous)



4

What are general linear models?

- You will see that GLMs w/ different variable types are just the “standard” tests you learn in STAT101 or see in publications!
- A lot of what you learned previously for linear regression (e.g., assumptions) applies here
- Main difference is learning how to interpret a slope w/ categorical variables
- GLM coefficients estimated w/ ordinary least squares

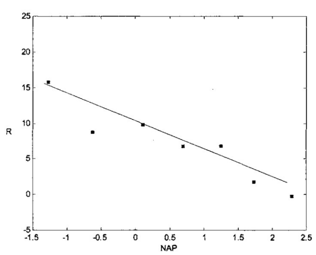
5

Generalized linear models (GLiM)

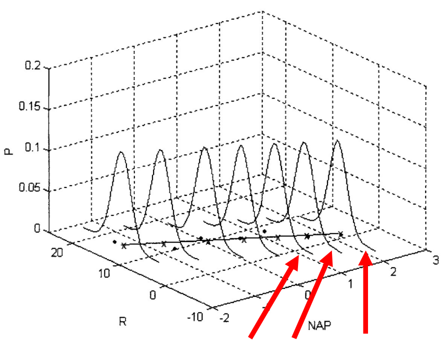
- GLMs assume normally distributed errors
  - Why you can use ordinary least squares
  - DV needs to be continuous
- GLiMs relax this assumption and allow errors to be non-normally distributed
  - E.g., logistic regression w/ binary DV & errors
- So GLM is a special version of GLiM, where errors are normal
- Coefficients estimated using maximum likelihood

6

An illustrative example



$Y \sim \text{Normal}(\beta_0 + \beta_1 X, \sigma^2)$



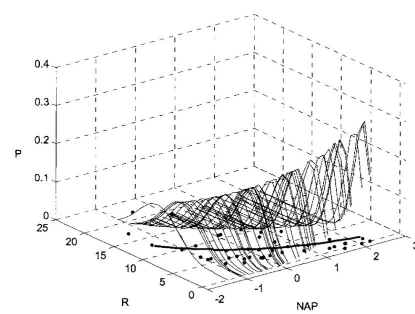
Zuur et al. 2007

7

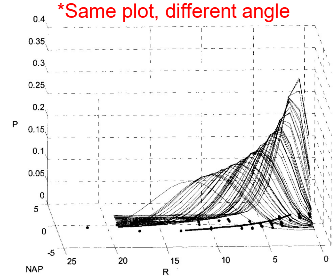
GLiM as a solution

Poisson regression → DV are count data, modeled as Poisson distributed (won’t cover this GLiM)

$Y \sim \text{Poisson}\ (\beta_0 + \beta_1 X)$



*Same plot, different angle

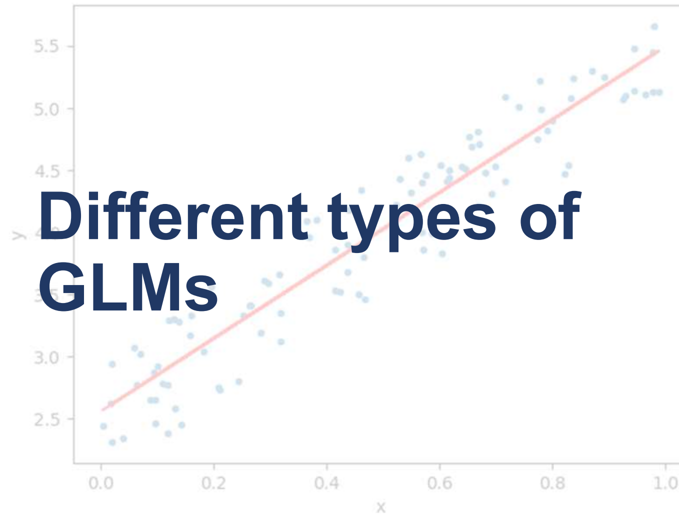


Zuur et al. 2007

8

Questions?

9



Different types of GLMs

10

Different types of GLMs/GLiMs

		Independent variable		
		Binomial	Multinomial	Continuous
Dependent variable	Binomial			
	Multinomial			
	Continuous			Regression

*Binomial and multinomial are both categorical variables w/ two and >2 categories, respectively

11

Two-sample t-test

Continuous DV ~ binomial IV

12

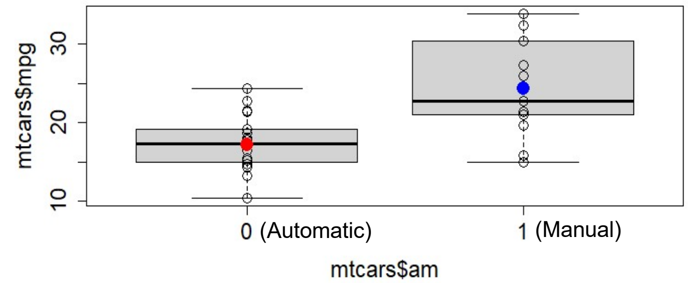
Different types of GLMs/GLiMs

		Independent variable		
Dependent variable		Binomial	Multinomial	Continuous
	Binomial			
	Multinomial			
	Continuous	t-test		Regression

*Binomial and multinomial are both categorical variables w/ two and >2 categories, respectively

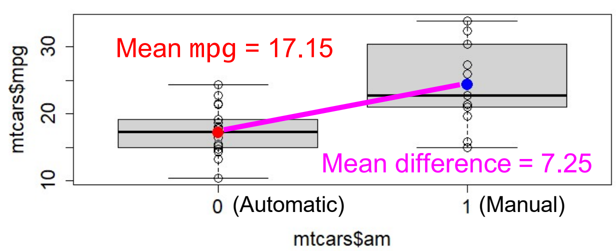
Two-sample t-test

- $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ Binomial IV (two levels)
- E.g., `mpg ~ am, data=mtcars`
  - am has two levels: 0 (automatic) & 1 (manual)



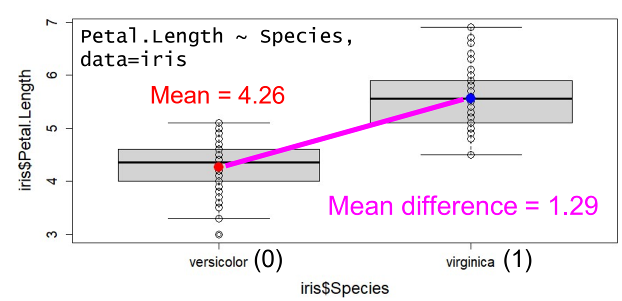
Two-sample t-test

- `mpg = 17.15 + 7.25am` (fitted linear model)
- Intercept is mean DV when IV = 0 (i.e., automatic), just like a normal intercept!
- Slope is change in mean DV as you go from 0 to 1 (i.e., manual), just like a normal slope!



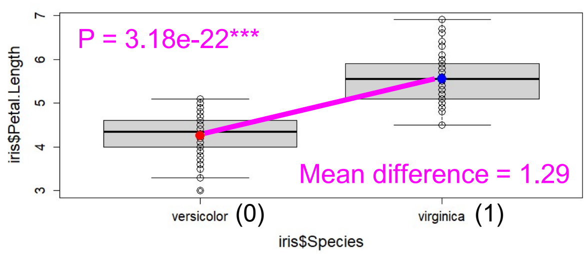
Dummy coding

- In general, your baseline level is coded as a 0, and the other is coded as a 1
- `Petal.length = 4.26 + 1.29Species`



Two-sample t-test

- **Used to test if two groups' means are significantly different**
- H_0: difference in groups' means = 0 → linear model slope = 0



Comparing `lm()` & `t.test()`

`lm()` (slope)	`t.test()`
• P = 3.18e-22***	• P = 3.18e-22***
• t = 12.60	• t = 12.60
• SE = 0.10	• SE = 0.10
• 95% CI = (1.09, 1.50)	• 95% CI = (1.09, 1.50)

t-test is **exactly** the same as a simple linear regression with a binomial IV!

Nonparametric tests

- Used when data w/in each group are not normally distributed (thus, errors are not normally distributed)
- BUT, central limit theorem ensures **mean or sum** is normally distributed if each group's sample size > 15 (general rule)
- Literally ranks DV and then performs test (e.g., like Spearman's)
- Due to less restrictive assumptions, less powerful than parametric counterpart (i.e., P-values are larger)

Mann-Whitney U test

- Nonparametric version of two-sample t-test
- Tests if two groups' *medians* are significantly different
- `wilcox.test(PL.virg, PL.vers)`
  - P = 9.13e-17 (compared w/ 3.18e-22 using t-test)

Questions?

21

ANOVA
(“analysis of variance”)

Continuous DV ~ multinomial IV

22

Different types of GLMs/GLiMs

		Independent variable		
		Binomial	Multinomial	Continuous
Dependent variable	Binomial			
	Multinomial			
	Continuous	t-test	ANOVA	Regression

*Binomial and multinomial are both categorical variables w/ two and >2 categories, respectively

23

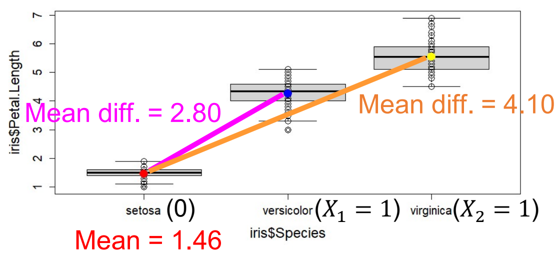
One-way ANOVA

- $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ Multinomial IV (>2 levels)
- E.g., `Petal.Length ~ Species`
  - `Species` has three levels: `setosa` (baseline), `versicolor`, and `virginica`

24

One-way ANOVA

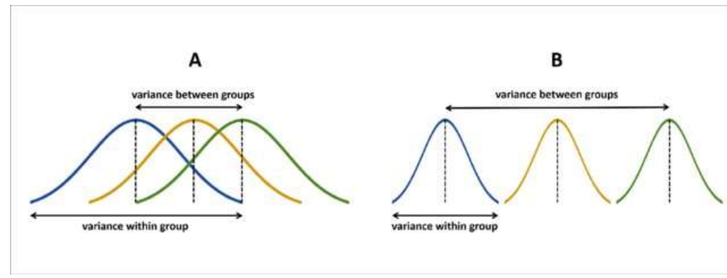
- So more accurately, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- `Petal.Length = 1.46 + 2.80versicolor + 4.10virginica`
- $N - 1$ estimated slopes (N = # levels in IV)



25

One-way ANOVA

- **Tests if groups' means are all equal** (w/ two groups, ANOVA is identical to a t-test)
- Calculates a *single* P-value using the F statistic (ratio of variance among groups to w/in groups)



26

Comparing `lm()` & `aov()`

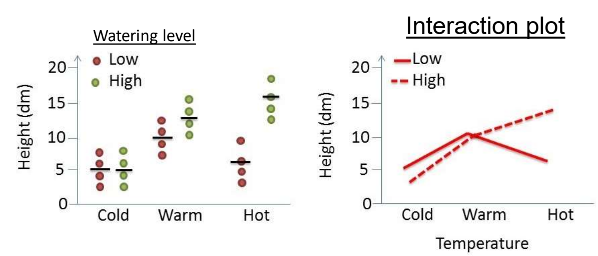
`lm()`
- F = 1180.2
- P = 2.86e-91***

`aov()`
- F = 1180.2
- P = 2.86e-91***

27

Two-way ANOVA

- Continuous DV ~ two multinomial IVs w/ an interaction term
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$



28

Kruskal-Wallis test

- Nonparametric version of ANOVA
- Tests if groups' *medians* are significantly different
- `kruskal.test(Petal.Length~Species)`
    - P = 4.80e-29 (compared w/ 2.86e-91 using ANOVA)

29

Questions?

30

ANCOVA
("analysis of covariance")

Continuous DV ~ categorical IV + continuous IV

31

Different types of GLMs/GLiMs

		Independent variable		
Dependent variable		Binomial	Multinomial	Continuous
	Binomial			
	Multinomial			
	Continuous	t-test	ANO ANCOVA	ression

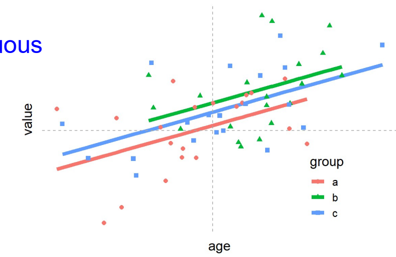
*Binomial and multinomial are both categorical variables w/ two and >2 categories, respectively

32

ANCOVA

- Combines regression w/ ANOVA
- **Used if regression intercept or slope varies as a function of levels w/in categorical IV**
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

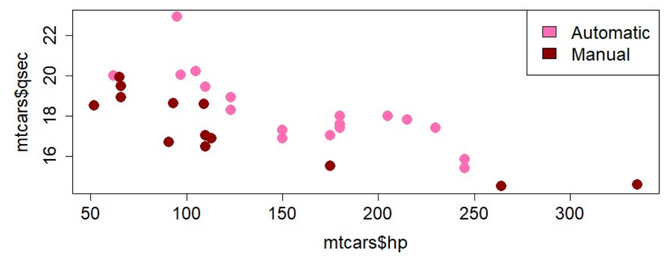
Categorical Continuous



33

Differing intercepts

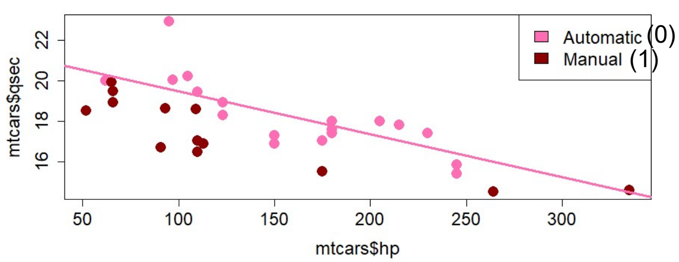
- **NO** interaction between IVs
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- E.g., `qsec ~ am + hp, data = mtcars`



34

Differing intercepts

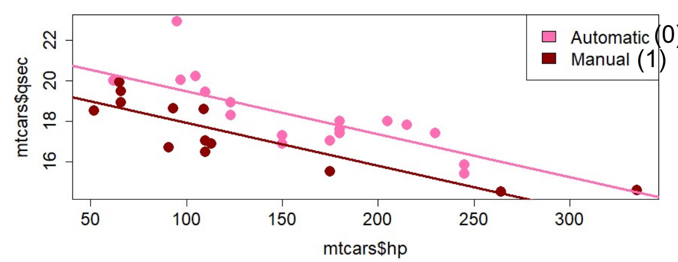
- qsec = 21.58 – 1.53am – 0.02hp
  = 21.58 – 1.53*0 – 0.02hp
  = 21.58 – 0.02hp



35

Differing intercepts

- qsec = 21.58 – 1.53am – 0.02hp
  = 21.58 – 1.53*1 – 0.02hp
  = 20.04 – 0.02hp



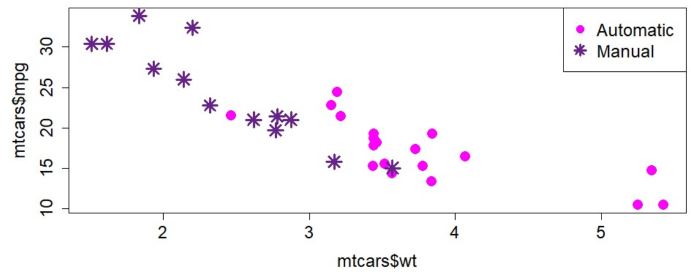
36

Interpreting coefficients

- `qsec = 21.58 – 1.53am – 0.02hp`
- `21.58` is estimated `qsec~hp` intercept for baseline level (i.e., automatic)
- `-1.53` is how much `qsec~hp` intercept changes going from automatic (0) to manual (1)
- Each additional level requires an additional coefficient (interpret from baseline level as in ANOVA)

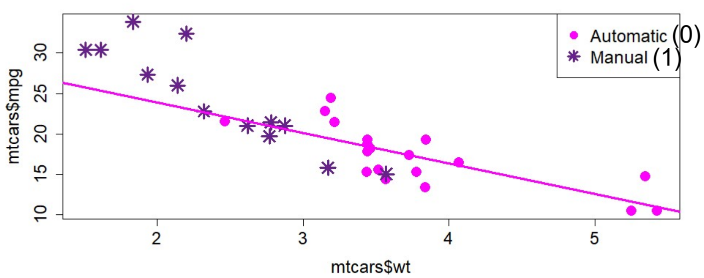
Differing intercepts & slopes

- **YES** interaction between IVs
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
- E.g., `mpg ~ am * wt, data = mtcars`



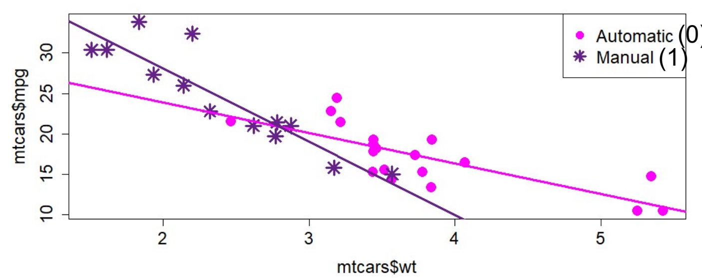
Differing intercepts & slopes

- `mpg = 31.42 + 14.88am – 3.79wt – 5.30am*wt`
- If `am = 0`, `hp = 31.42 – 3.79wt`



Differing intercepts & slopes

- `mpg = 31.42 + 14.88am – 3.79wt – 5.30am*wt`
- If `am = 1`, `mpg = 46.29 – 9.08wt`



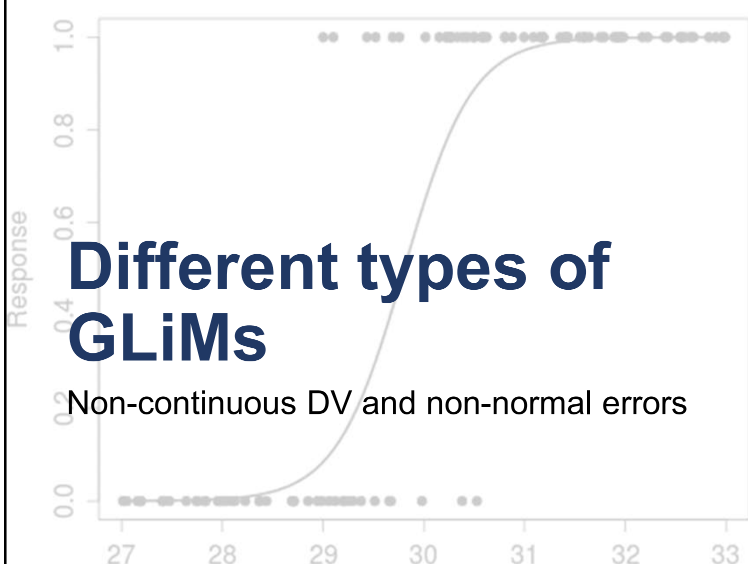
Interpreting coefficients

- `mpg = 31.42 + 14.88am – 3.79wt – 5.30am*wt`
- `31.42` is estimated `mpg~wt` intercept for baseline level (i.e., automatic)
- `-3.79` is estimated `mpg~wt` slope for baseline level (i.e., automatic)
- `14.88` is how much `mpg~wt` intercept changes going from automatic (0) to manual (1)
- `-5.30` is how much `mpg~wt` slope changes going from automatic (0) to manual (1)

41

Questions?

42



Different types of GLiMs

Non-continuous DV and non-normal errors

43

Generalized linear models

- Thus far, we have covered GLMs (where DV is continuous & errors are normally distributed) w/ IVs of different data types
- Now we move onto GLiMs, where the DV's data type changes (thus causing non-normal errors)

44

Logistic regression

Binomial DV ~ continuous IV

Different types of GLMs/GLiMs

Dependent variable \ Independent variable	Binomial	Multinomial	Continuous
Binomial			Logistic regression
Multinomial			
Continuous	t-test	ANOVA	Regression

ANCOVA

*Binomial and multinomial are both categorical variables w/ two and >2 categories, respectively

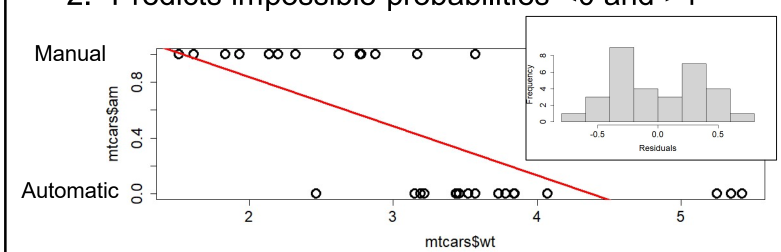
Logistic regression

- Binomial DV ~ one or more IVs (usually continuous but can be categorical)
- What are some examples of a binomial DV in your field?
- **Used to assess probability of belonging to non-baseline level as a function of IVs**
- E.g., `am ~ wt, data = mtcars`
  - Probability car is manual `(am=1)` as `wt` increases

`am ~ wt, data = mtcars`

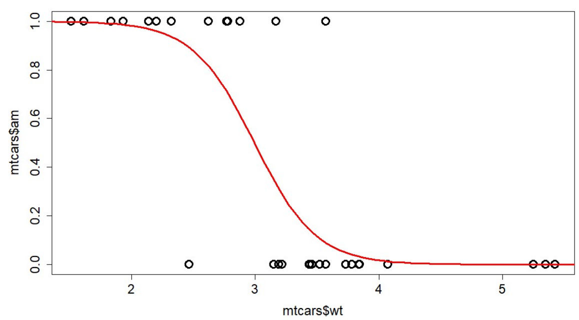
- More likely to be manual if car is lighter
- But linear regression model is terrible!

1. Relationship is not linear; errors not normal
2. Predicts impossible probabilities <0 and >1



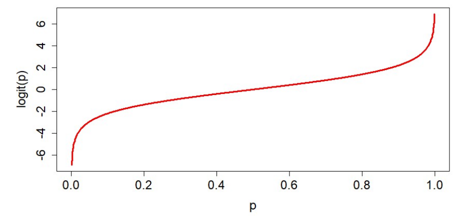
`am ~ wt, data = mtcars`

- A logistic function is better
- Minimum probability is zero, maximum is one



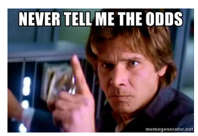
Logit transformation

- Logistic regression uses a *logit transformation* to convert logistic curve → straight line, so DV probabilities can be modeled w/ linear model
- $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$



- logit(p) goes from negative infinity to infinity
- All done under the hood in R

Interpreting coefficients



- First, a primer on odds
- If p is the probability of something happening, odds are $\frac{p}{1-p}$
- E.g., if probability of drawing a card w/ clubs is 0.25, odds are $0.25/0.75 = 0.33$
  - You're three times less likely to get clubs
- E.g., if probability of rolling a 1, 2, 3, or 4 w/ a die is 0.66, odds are $0.66/0.33 = 2$
  - You're twice as likely to roll these numbers
- **Odds < 1 means event less likely to happen; odds > 1 means event more likely to happen**

Interpreting coefficients

- `am ~ wt, data = mtcars`
- $\log\left(\frac{p}{1-p}\right) = 12.04 - 4.02\mathrm{wt}$
- exp(intercept) is odds car will be manual when `wt`=0
  - $\exp(12.04) = 169{,}397$
- exp(slope) is proportional change in odds car will be manual when `wt` increases by one
  - $\exp(-4.02) = 0.02$ → odds decrease by 98%!

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Demonstrating this algebraically

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \rightarrow \frac{p}{1-p} = \exp(\beta_0 + \beta_1 x) = \exp(\beta_0)\exp(\beta_1 x)$

Intercept

$\frac{p}{1-p} = \exp(\beta_0)\exp(\beta_1 \cdot 0) = \exp(\beta_0)$

$\frac{p}{1-p} = \exp(\beta_0)$

Compare

Slope

$\frac{p}{1-p} = \exp(\beta_0)\exp(\beta_1 \cdot (x+1)) = \exp(\beta_0)\exp(\beta_1 x + \beta_1) =$
$\exp(\beta_0)\exp(\beta_1 x)\underline{\exp(\beta_1)}$

Thus, increasing x by 1 increases odds by the multiplier, $\exp(\beta_1)$

53

Questions?

54

Multinomial logistic regression

Multinomial DV ~ continuous IV

55

Different types of GLMs/GLiMs

		Independent variable		
Dependent variable		Binomial	Multinomial	Continuous
	Binomial			Logistic regression
	Multinomial			Multinomial regression
	Continuous	t-test	ANOVA	Regression

ANCOVA

*Binomial and multinomial are both categorical variables w/ two and >2 categories, respectively

56

Multinomial logistic regression

- Multinomial DV ~ one or more IVs (usually continuous but can be categorical)
- **Used to assess odds of belonging to <u>EACH</u> non-baseline level as a function of IVs**
- Coefficients interpreted in same way as in logistic regression

https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

57

Questions?

58

Chi-squared test

Categorical DV ~ categorical IV

59

Different types of GLMs/GLiMs

		Independent variable		
Dependent variable		<u>Binomial</u>	<u>Multinomial</u>	<u>Continuous</u>
	<u>Binomial</u>	Chi-squared	Chi-squared	Logistic regression
	<u>Multinomial</u>	Chi-squared	Chi-squared	Multinomial regression
	<u>Continuous</u>	t-test	ANOVA	Regression

ANCOVA

*Binomial and multinomial are both categorical variables w/ two and >2 categories, respectively

60

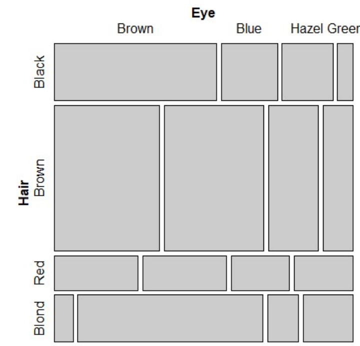
Pearson's chi-squared test

- One categorical DV ~ one categorical IV
- Data presented as a **contingency table** (AKA **crosstab**)

	Eye color			
Hair color	Brown	Blue	Hazel	Green
Black	32	11	10	3
Brown	53	50	25	15
Red	10	10	7	7
Blond	3	30	5	8

Mosaic plot

- Cool way to visualize a contingency table



Pearson's chi-squared test

- Categorical DV ~ one categorical IV
- Data presented as a **contingency table** (AKA **crosstab**)
- **Tests H_0 of whether two categorical variables are independent of each other**
  - e.g., if certain hair colors are NOT associated w/ certain eye colors
- Independence operationalized as cell frequencies that are proportional to column & row totals

Pearson's chi-squared test

- H_0 expected = (row total x column total) / grand total
- χ^2 test statistic: $\sum_{\text{al cells}} \frac{(\text{Observed}-\text{Expect}\quad)^2}{\text{Expected}}$
- χ^2 statistic used to get P-value

	Eye color (56 x 33) / 279 = 6.6				
Hair color	Brown	Blue	Hazel	Green	**Total**
Black	32	11	10	3	**56**
Brown	53	50	25	15	**143**
Red	10	10	7	7	**34**
Blond	3	30	5	8	**46**
Total	**98**	**101**	**47**	**33**	**279**

Pearson’s chi-squared test

- Also a log-linear model (a generalized linear model for DV of counts): <u>frequencies ~ IV * DV</u>

	Freq	hair_color	eye_color
1	32	Black	Brown
2	53	Brown	Brown
3	10	Red	Brown
4	3	Blond	Brown
5	11	Black	Blue
6	50	Brown	Blue
7	10	Red	Blue
8	30	Blond	Blue

65

Pearson’s chi-squared test

- Also a log-linear model (a generalized linear model for DV of counts): <u>frequencies ~ IV * DV</u>
- The interaction term is what is tested in a chi-squared test

`chisq.test()`

- χ^2 = 41.28
- P = 4.45e-6

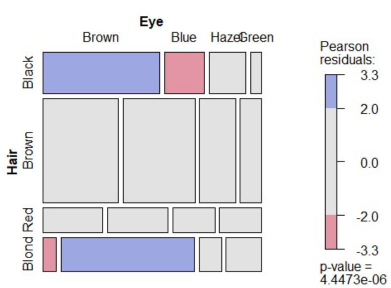
`log-linear`

- χ^2 = 41.28
- P = 4.45e-6

66

Significance driven by:

- Overrepresentation of <u>black hair/brown eyes</u> and <u>blond hair/blue eyes</u>
- Underrepresentation of <u>black hair/blue eyes</u> and <u>blonde hair/brown eyes</u>



Pearson residuals = $\dfrac{(\text{observed} - \text{expected})}{\sqrt{\text{expected}}}$

67

Questions?

68

Summary: GLMs/GLiMs

		Independent variable		
Dependent variable		Binomial	Multinomial	Continuous
	Binomial	Chi-squared	Chi-squared	Logistic regression
	Multinomial	Chi-squared	Chi-squared	Multinomial regression
	Continuous	t-test	ANOVA	Regression

ANCOVA

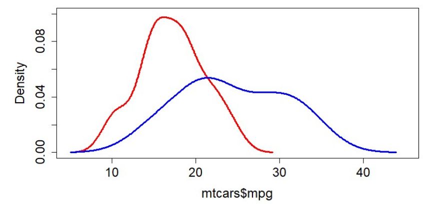
*Binomial and multinomial are both categorical variables w/ two and >2 categories, respectively

Summary

- t-tests, ANOVAs, & chi-squared tests emphasize P-values, so I am not a fan
- Presenting means and SD of each group & plots are more informative (to me)

Which do you think is more informative?

P = 0.001 vs.



Statistics vignette

- Are declining SAT scores bad for the country?

SAT®

Steve Wang

Decline in average SAT reading scores

- 1972: 530
- 2011: 497

Average SAT scores by state

1. Illinois
2. Minnesota
3. Iowa
4. Wisconsin
5. Missouri
6. Michigan
7. North Dakota
8. Kansas
9. Nebraska
10. South Dakota

27. Massachusetts

30. Vermont
31. Connecticut
33. California

42. New York

Missing some information…

- What percentage of high schoolers take the SAT in each state?

Average SAT scores by state

1. Illinois (5%)
2. Minnesota (7%)
3. Iowa (3%)
4. Wisconsin (5%)
5. Missouri (5%)
6. Michigan (5%)
7. North Dakota (3%)
8. Kansas (6%)
9. Nebraska (5%)
10. South Dakota (4%)

27. Massachusetts (89%)

30. Vermont (67%)
31. Connecticut (87%)
33. California (53%)

42. New York (89%)

Trends through time

Since 1991, number of test takers has gone up 59%

From 1950 to 2011, proportion w/ four-year degree: 6% to 30%