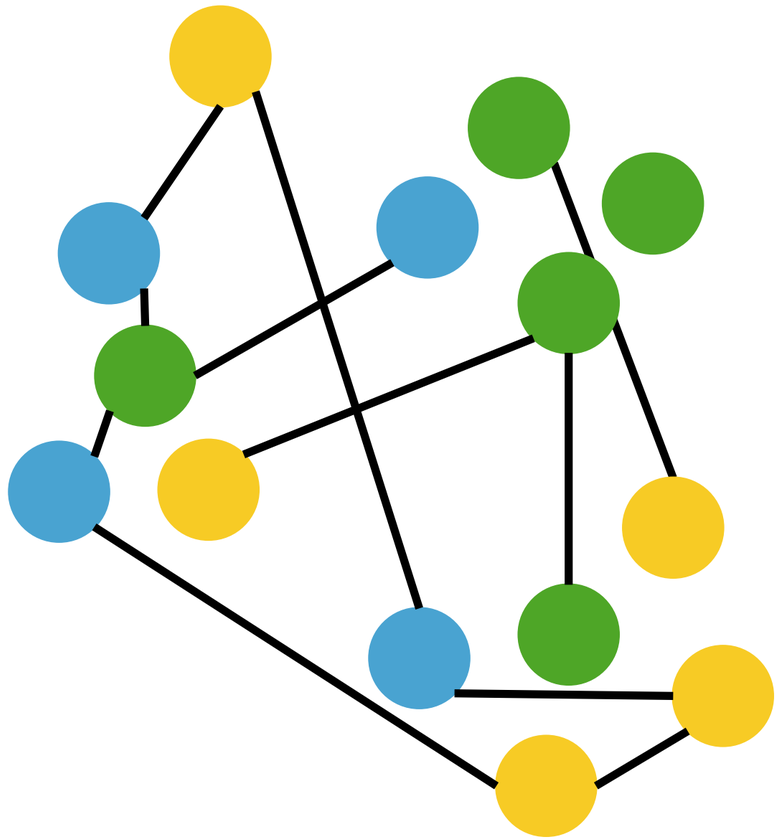
Numerical Problems

Announcements

- Really appreciate your feedback on ChatGPT
  - For those of you who expressed concerns about some students using it, I have given 0's on problems where I identified its blatant use
    - Still looking through exams for any other signs
  - Unfortunately, we're in an in-between phase while SPH works on developing school policies on ChatGPT's use
  - One thing I want to say: whether you think it's a good tool or not, there is certain expertise you should come out of this class with
    - In this class we are building nodes of knowledge and connecting those nodes
- I'm playing catch up on all communication this week!



Project information

- Education in screening data
  - Sorry for the back and forth on this
  - Please include education in your analysis
- Missing data
  - Complete case analysis for our class
    - "drop_na()" in R should help
      - But make sure you drop NA's in data frame that includes only your selected variables
- Presentations
  - Given by the group
  - Ideally want each person to talk for some time
  - Will give more details and update project instructions

github
Rstudio in
One Drive
(Share point?)

Google drive

proposal
due Mon
29th

Midterm

- Please come talk to me if you have questions
  - Or email!
- There is still one student who has not taken the test, so we will not discuss answers in class
- A couple questions that were challenging for majority of the class
- Midterm feedback (mentioned in syllabus) will be sent out soon
  - Will help boost grades
- Overall, very impressed with what you accomplished!

Homeworks

- Homework 3
  - Redo due Friday, May 26, at 11pm
- Homework 4
  - Graded and Released
  - Redo due Friday, June 2, at 11pm
- Homework 5
  - Due Thursday, May 25, at 11pm

HW 6 optional?

Numerical Problems

"Successful modeling of a complex data set is **part science**, **part statistical methods**, and **part experience and common sense**."

Hosmer, Lemeshow, and Sturdivant Textbook, pg. 101

Class 13 Learning Objectives

1. Identify and troubleshoot logistic regression analysis when there are zero observations for the cross section of the outcome and a predictor

2. Identify and troubleshoot logistic regression analysis when there is complete separation between the two outcome groups

3. Identify and troubleshoot logistic regression analysis when there is collinearity between variables

Numerical Problems

- Issues that may cause numerical problems:

1. Zero in a contingency table
2. Complete separation
3. Collinearity

- All may cause large estimated coefficients and/or large estimated standard errors

Class 13 Learning Objectives

1. Identify and troubleshoot logistic regression analysis when there are zero observations for the cross section of the outcome and a predictor

2. Identify and troubleshoot logistic regression analysis when there is complete separation between the two outcome groups

3. Identify and troubleshoot logistic regression analysis when there is collinearity between variables

Zero cell in a contingency table

- If no observations at any intersection of the covariate and outcome
- Zero cell in a contingency table should be detected in descriptive statistical analysis stage
- Example of one covariate with outcome:

Outcome	Covariate (x)			Total
	1	2	3	
1	7	12	20	39
0	13	8	0	21
Total	20	20	20	60

Zero cell in a contingency table

- Example of one covariate with outcome:

```
ex1_glm = glm(outcome ~ x, data = ex1, family = binomial())
summary(ex1_glm)
```

```
##
## Call:
## glm(formula = outcome ~ x, family = binomial(), data = ex1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.35373  -0.92821   0.00008   1.01077   1.44901
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.6190     0.4688  -1.320    0.187
## xThree        20.1851  2404.6704   0.008    0.993
## xTwo           1.0245     0.6543   1.566    0.117
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 77.694  on 59  degrees of freedom
## Residual deviance: 53.819  on 57  degrees of freedom
```

Outcome	Covariate (x)			Total
	1	2	3	
1	7	12	20	39
0	13	8	0	21
Total	20	20	20	60

Zero cell in a contingency table

- Example of one covariate with outcome:

```
ex1_glm = glm(outcome ~ x, data = ex1, family = binomial())
summary(ex1_glm)
```

```
##
## Call:
## glm(formula = outcome ~ x, family = binomial(), data = ex1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.35373  -0.92821   0.00008   1.01077   1.44901
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.6190     0.4688  -1.320    0.187
## xThree        20.1851  2404.6704   0.008    0.993
## xTwo           1.0245     0.6543   1.566    0.117
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 77.694  on 59  degrees of freedom
## Residual deviance: 53.818  on 57  degrees of freedom
```

Outcome	Covariate (x)			Total
	1	2	3	
1	7.5	12	20	39
0	13.5	8	0.5	21
Total	20	20	20	60

12 20
12 0

Coefficient estimate is large and standard error of coefficient is large!

Zero cell in a contingency table

- Ways to address the issue (zero cell in a contingency table):
  - Add one-half to each of the cell counts
    - Technically works, but not the best option
    - May work for simple logistic regression
    - Rarely useful with a more complex analysis
  - Collapse the categories to remove the 0 cells
    - We could collapse groups 2 and 3 together if it makes clinical sense
  - Remove the category with 0 cells
    - This would mean we reduce the total sample size as well
  - If the variable is in ordinal scale, treat it as continuous
    - But check the linearity!!

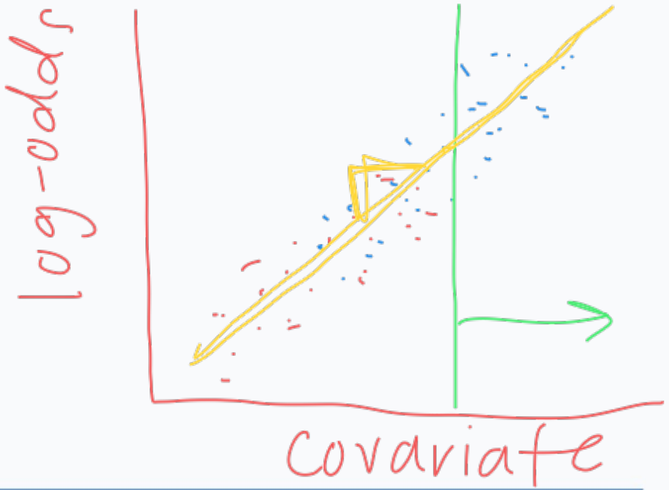
Poll Everywhere Question 1

Can the zero cell problem apply to a continuous covariate/predictor of an outcome?

https://www.polleverywhere.com/multiple_choice_polls/QP16XBShi95619wyNnSOn

Respond at **PollEv.com/nickywakim275**

Can the zero cell problem apply to a continuous covariate/predictor of an outcome?



Powered by Poll Everywhere

Zero cell in a contingency table: interaction

- Often when we add in interactions, we don't look at the contingency table first
- We are looking at the results of a fit model with the interaction
- Interaction term for x and third z group looks suspicious

x - binary
z - 3 groups

$$\text{logit}(\pi(X_i, Z_i)) = \beta_0 + \beta_1 X_i + \beta_2 I(Z_i=2) + \beta_3 I(Z_i=3) + \beta_4 X_i \cdot I(Z_i=2) + \beta_5 X_i I(Z_i=3)$$

Table 4.36 Results of Fitting Logistic Regression Models to the Data in Table 4.35

Model	1		2	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.
x	2.77	0.72	1.39	1.01
z_2	1.19	0.81	0.29	1.14
z_3	2.04	0.89	0.00	1.37
$x \times z\_2$			1.32	1.51
$x \times z\_3$			11.54	50.22
Constant	−2.32	0.77	−1.39	0.79

Zero cell in a contingency table: interaction

- Contingency table used to detect zero cell issue when interaction term is included in the logistic regression model
- Contingency table: interaction of z and x

Table 4.35 Stratified 2 by 2 Contingency Tables with a Zero Cell Count Within One Stratum

Stratum (z)		1		2		3	
Outcome /	x	1	0	1	0	1	0
1		5	2	10	2	15	1
0		5	8	2	6	0	4
Total		10	10	12	8	15	5
$\widehat{\text{OR}}$			4		15		inf

Zero cell in a contingency table

- Ways to address the issue (zero cell in a contingency table):
  - Add one-half to each of the cell counts
  - Collapse the categories to remove the 0 cells
    - We could collapse the second and third z groups
    - Not an option to collapse x
  - Remove the category with 0 cells
    - Remove z = 3, but then we lose ~~15~~ 20 observations for x as well
- If the variable is in ordinal scale, treat it as continuous
  - Could treat z as continuous

Class 13 Learning Objectives

1. Identify and troubleshoot logistic regression analysis when there are zero observations for the cross section of the outcome and a predictor

2. Identify and troubleshoot logistic regression analysis when there is complete separation between the two outcome groups

3. Identify and troubleshoot logistic regression analysis when there is collinearity between variables

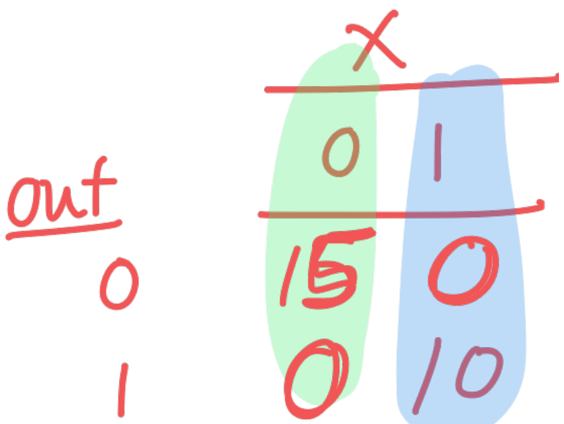
gladder for linearity?

Complete Separation

- Complete separation: occurs when a collection of the covariates completely separates the outcome groups
  - Example: Outcome is “gets senior discount at iHop” and the only covariate you measure is age
    - Age will completely separate the outcome
    - No overlap in distribution of covariates between two outcome groups

- Problem: the maximum likelihood estimates do not exist
  - Likelihood function is monotone
  - In order to have finite maximum likelihood estimates we must have some overlap in the distribution of the covariates in the model

Poll Everywhere Question 2



True or False? Complete separation of an outcome with a binary covariate/predictor can be viewed as a zero cell problem.

https://www.polleverywhere.com/multiple_choice_polls/sHOA1Tp58QD8X3igQUZyH

Respond at **PollEv.com/nickywakim275**

True or False? Complete separation of an outcome with a binary covariate/predictor can be viewed as a zero cell problem.

Powered by Poll Everywhere

Complete Separation: example

```
y = c(0,0,0,0,1,1,1,1)
x1 = c(1,2,3,3,5,6,10,11)
x2 = c(3,2,-1,-1,2,4,1,0)
ex3 = data.frame(outcome = y, x1 = x1, x2= x2)
ex3
```

```
##   outcome x1 x2
## 1       0  1  3
## 2       0  2  2
## 3       0  3 -1
## 4       0  3 -1
## 5       1  5  2
## 6       1  6  4
## 7       1 10  1
## 8       1 11  0
```

```
m1 = glm(y ~ x1 + x2, family=binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Complete Separation: example

```
summary(m1)
```

```
##
## Call:
## glm(formula = y ~ x1 + x2, family = binomial)
##
## Deviance Residuals:
##          1           2           3           4           5           6
## -2.110e-08  -1.404e-05  -2.522e-06  -2.522e-06   1.564e-05   2.110e-08
##          7           8
##  2.110e-08   2.110e-08
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -66.098 183471.722   0.000        1
## x1             15.288  27362.843   0.001        1
## x2              6.241  81543.720   0.000        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.1090e+01  on 7  degrees of freedom
## Residual deviance: 4.5454e-10  on 5  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 24
```

Complete Separation: example

```
summary(m1)
```

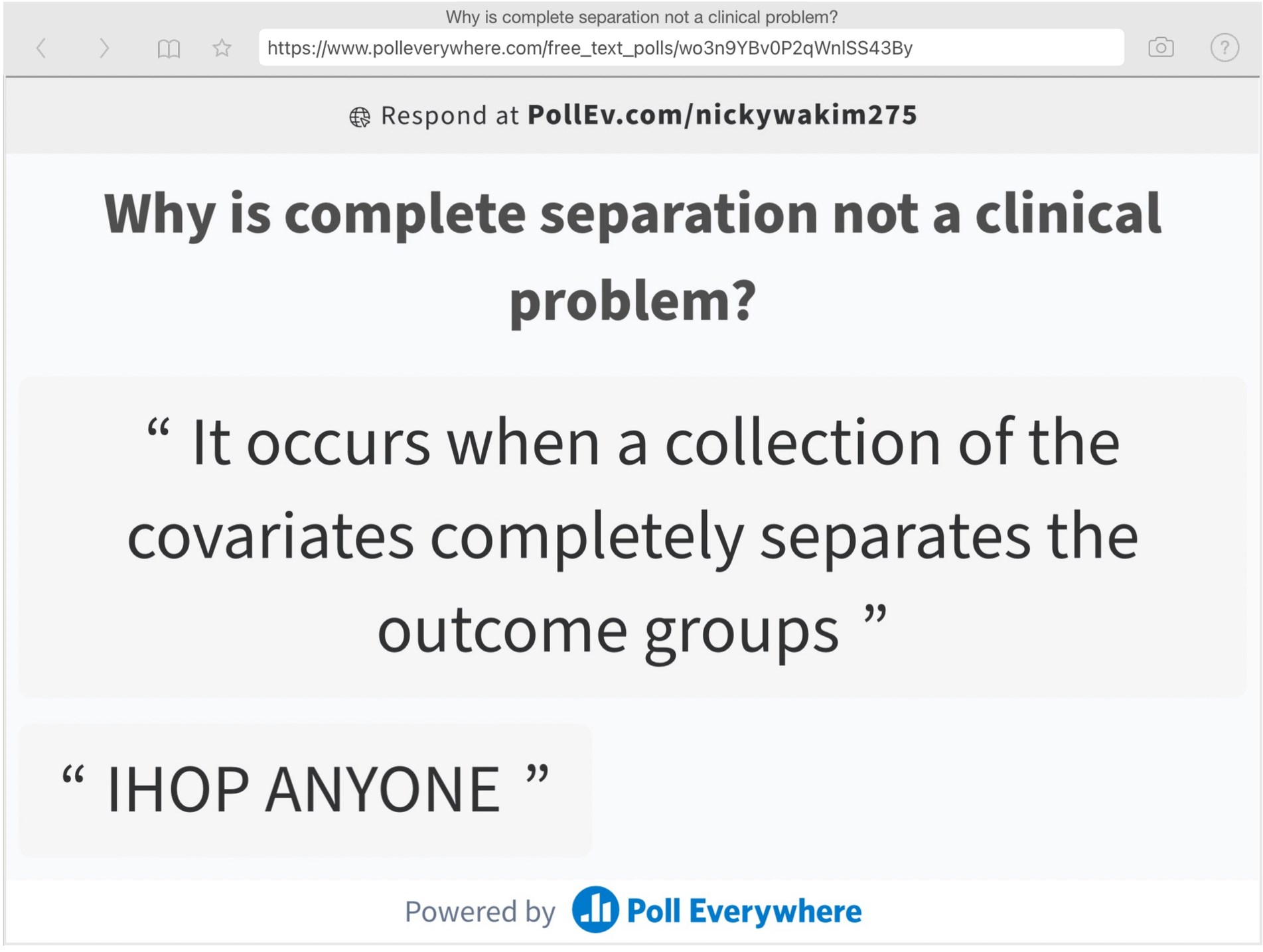
```
##
## Call:
## glm(formula = y ~ x1 + x2, family = binomial)
##
## Deviance Residuals:
##          1           2           3           4           5           6
## -2.110e-08  -1.404e-05  -2.522e-06  -2.522e-06   1.564e-05   2.110e-08
##          7           8
##  2.110e-08   2.110e-08
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -66.098 183471.722   0.000        1
## x1             15.288  27362.843   0.001        1
## x2              6.241  81543.720   0.000        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.1090e+01  on 7  degrees of freedom
## Residual deviance: 4.5454e-10  on 5  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 24
```

Coefficient estimate is large and standard error of coefficient is large!

Complete Separation

- The occurrence of complete separation in practice depends on
  - Sample size
  - Number of subjects with the outcome present
  - Number of variables included in the model
- Example: 25 observations and only 5 have “success” outcome
  - 1 variable in model may not lead to complete separation
  - More variables = more dimensions that can completely separate the observations
- In most cases, the occurrence of complete separation is not bad for clinical importance
  - But rather a numerical coincidence that causing problem for model fitting

Poll Everywhere Question 3



Complete Separation

- Ways to address the issue:
  - Collapse categorical variables in a meaningful way
    - Easiest and best if stat methods are restricted (common for collaborations)
  - Exclude x1 from the model
    - Not ideal because this could lead to biased estimates for the other predicted variables in the model
- Firth logistic regression
  - Uses penalized likelihood estimation method
  - Basically takes the likelihood (that has no maximum) and adds a penalty that makes the MLE estimatable

Complete Separation

- Firth logistic regression

```
library(logistf)
m1_f = logistf(y ~ x1 + x2, family=binomial)
summary(m1_f)
```

replaces glm()

data = ex3)

```
## logistf(formula = y ~ x1 + x2, family = binomial)
##
## Model fitted by Penalized ML
## Coefficients:
##                   coef  se(coef)   lower 0.95 upper 0.95     Chisq          p
## (Intercept) -2.9748898 1.7244237 -15.47721665 -0.1208883 4.2179522 0.03999841
## x1           0.4908484 0.2745754   0.05268216  2.1275832 5.0225056 0.02501994
## x2           0.4313732 0.4988396  -0.65793078  4.4758930 0.7807099 0.37692411
##              method
## (Intercept)       2
## x1                2
## x2                2
##
## Method: 1-Wald, 2-Profile penalized log-likelihood, 3-None
##
## Likelihood ratio test=5.505687 on 2 df, p=0.06374636, n=8
## Wald test = 3.624899 on 2 df, p = 0.1632538
```

Class 13 Learning Objectives

1. Identify and troubleshoot logistic regression analysis when there are zero observations for the cross section of the outcome and a predictor

2. Identify and troubleshoot logistic regression analysis when there is complete separation between the two outcome groups

3. Identify and troubleshoot logistic regression analysis when there is collinearity between variables

Collinearity

- Collinearity happens when one or more of the covariates in a model can be predicted from other covariates in the same model

- This will cause unreliable coefficient estimates for some covariates in logistic regression, as in an ordinary linear regression

- Looking at correlations among pairs of variables is helpful but not enough to identify collinearity problem
  - Because collinearity problems may involve relationships among more than two covariates

Collinearity

- Table below is a simulated data with $x_1 \sim N(0,1), x_2 = x_1 + U(0,0.1), x_3 = 1 + U(0, 0.01)$
- Therefore, x_1 and x_2 are highly correlated, and x_3 is nearly collinear with the constant term.

Table 4.38 Data Displaying Near Collinearity Among the Independent Variables and Constant

Subject	x_1	x_2	x_3	y
1	0.225	0.231	1.026	0
2	0.487	0.489	1.022	1
3	−1.080	−1.070	1.074	0
4	−0.870	−0.870	1.091	0
5	−0.580	−0.570	1.095	0
6	−0.640	−0.640	1.010	0
7	1.614	1.619	1.087	0
8	0.352	0.355	1.095	1
9	−1.025	−1.018	1.008	0
10	0.929	0.937	1.057	1

Collinearity

- Four logistic regression models using data in the previous slide
- Consequence of collinearity: large coefficient estimates and/or standard errors

Table 4.39 Estimated Coefficients and Standard Errors from Fitting Logistic Regression Models to the Data in Table 4.38

Var.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
x_1	1.4	1.0	104.2	256.2			79.8	272.6
x_2			−103.4	256.0			−78.3	272.5
x_3					1.8	20.0	−11.1	206.6
Cons.	−1.0	0.8	−0.3	1.3	−2.7	21.1	11.4	27.8

Collinearity

- Four logistic regression models using data in the previous slide
- Consequence of collinearity: large coefficient estimates and/or standard errors

Table 4.39 Estimated Coefficients and Standard Errors from Fitting Logistic Regression Models to the Data in Table 4.38

Var.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
x_1	1.4	1.0	104.2	256.2			79.8	272.6
x_2			−103.4	256.0			−78.3	272.5
x_3					1.8	20.0	−11.1	206.6
Cons.	−1.0	0.8	−0.3	1.3	−2.7	21.1	11.4	27.8
	Model 1: x1 only		Model 2: x1 and x2		Model 3: x3 only		Model 4: x1, x2, x3	

Collinearity

- Four logistic regression models using data in the previous slide
- Consequence of collinearity: large coefficient estimates and/or standard errors

Table 4.39 Estimated Coefficients and Standard Errors from Fitting Logistic Regression Models to the Data in Table 4.38

Var.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
x_1	1.4	1.0	104.2	256.2			79.8	272.6
x_2			−103.4	256.0			−78.3	272.5
x_3					1.8	20.0	−11.1	206.6
Cons.	−1.0	0.8	−0.3	1.3	−2.7	21.1	11.4	27.8

Model 1: x1 only

Model 2: x1 and x2

Large coefficients and SE

Model 3: x3 only

Large SE

Model 4: x1, x2, x3

Large coefficients and SE

Collinearity

- Collinearity only involves the covariates
  - No specific issues to logistic regression (vs. linear regression)
  - Techniques from 512/612 work well for logistic regression model

- In more complicated dataset/analysis, we may not be able to detect collinearity using the coefficient estimates/SE

- **Variance inflation factor (VIF) approach:** well-known approach to detect collinearity

Variance Inflation Factor (VIF) Approach

- Computed by regressing each variable on all the other explanatory variables
  - For example: $E(x_1|x_2, x_3, \ldots) = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$
- Calculate the coefficient of determination, R^2
  - Proportion of the variation in x_1 that is predicted from $x_2, x_3, \ldots$

$$VIF = \frac{1}{1 - R^2}$$

- Each covariate has its own VIF computed
- Get worried for collinearity if VIF > 10
- Sometimes VIF approach may miss serious collinearity
  - Same collinearity we wish to detect using VIF can cause numerical problems in reliably estimating R^2

GLOW Study: Running example for class

- Variables in the study
- Fracture (last variable in table) is our outcome

Table 1.7 Code Sheet for Variables in the GLOW Study

Variable	Description	Codes/Values	Name
1	Identification code	1–*n*	SUB_ID
2	Study site	1–6	SITE_ID
3	Physician ID code	128 unique codes	PHY_ID
4	History of prior fracture	1 = Yes 0 = No	PRIORFRAC
5	Age at enrollment	Years	AGE
6	Weight at enrollment	Kilograms	WEIGHT
7	Height at enrollment	Centimeters	HEIGHT
8	Body mass index	kg/m^2	BMI
9	Menopause before age 45	1 = Yes 0 = No	PREMENO
10	Mother had hip fracture	1 = Yes 0 = No	MOMFRAC
11	Arms are needed to stand from a chair	1 = Yes 0 = No	ARMASSIST
12	Former or current smoker	1 = Yes 0 = No	SMOKE
13	Self-reported risk of fracture	1 = Less than others of the same age 2 = Same as others of the same age 3 = Greater than others of the same age	RATERISK
14	Fracture risk score	Composite risk score[a]	FRACSCORE
15	Any fracture in first year	1 = Yes 0 = No	FRACTURE

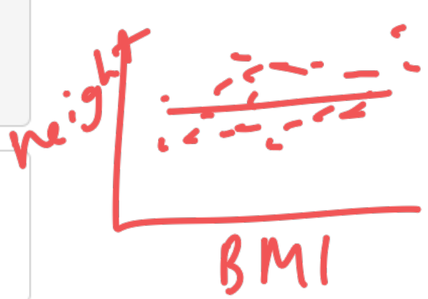
GLOW study: height, weight, BMI

```
main_eff = glm(fracture ~ age + height + priorfrac + momfrac +
                 armassist + raterisk2,
                 data = glow2, family = binomial)

glow_model = glm(fracture ~ age + height + priorfrac + momfrac +
                   armassist + raterisk2 + weight + bmi,
                 data = glow2, family = binomial)
```

```
library(car)
vif(main_eff)
```

```
##       age    height priorfrac   momfrac armassist raterisk2
##  1.169168  1.067010  1.118853  1.021981  1.103161  1.067098
```




```
vif(glow_model)
```

```
##        age     height  priorfrac   momfrac  armassist  raterisk2     weight
##   1.368259  19.561368   1.134230  1.028988   1.300983   1.120429 166.214541
##        bmi
## 152.406849
```

Collinearity

- Ways to address the issue:
  - Exclude the redundant variable from the model
    - In GLOW study, saw that height is okay without weight and BMI
  - Scaling and centering variables
    - When you have transformed a continuous variable
- Other modeling approach (outside scope of this class)
  - Ridge regression
  - Principle component analysis

Poll Everywhere Question 4

Collinearity

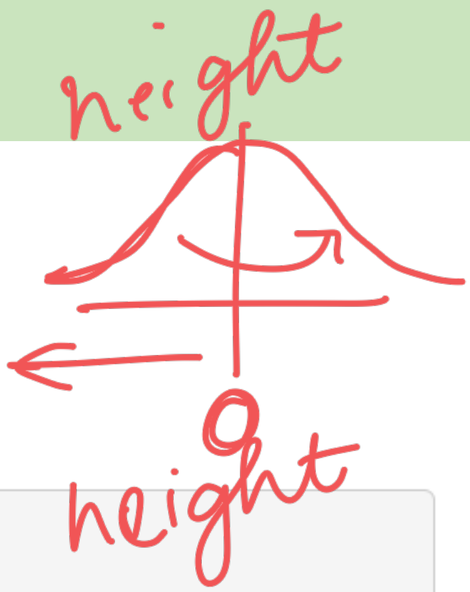
- Scaling and centering variables
- Let's say we found that height is nonlinearly related to fracture, so include a squared height term

```
glow3 = glow2 %>% mutate(height_sq = height^2)
height2 = glm(fracture ~ age + height + priorfrac + momfrac +
                armassist + raterisk2 + height_sq,
                data = glow3, family = binomial)
# summary(height2)
vif(height2)
```

```
##        age     height  priorfrac    momfrac  armassist  raterisk2  height_sq
##   1.173304 571.542009   1.120032   1.023053   1.101958   1.069178 571.172305
```

Collinearity

- Scaling and centering variables
- Centering height will reduce collinearity while still helping the nonlinear problem!




```
glow4 = glow2 %>% mutate(height_c = (height - mean(height)),
                         height_c_sq = height_c^2)
height_c_2 = glm(fracture ~ age + height_c + priorfrac + momfrac +
                 armassist + raterisk2 + height_c_sq,
                 data = glow4, family = binomial)
# summary(height_c_2)
vif(height_c_2)
```

```
##          age     height_c   priorfrac     momfrac   armassist    raterisk2
##     1.173304     1.072721     1.120032     1.023053     1.101958     1.069178
## height_c_sq
##     1.006067
```

Class 13 Learning Objectives

1. Identify and troubleshoot logistic regression analysis when there are zero observations for the cross section of the outcome and a predictor

2. Identify and troubleshoot logistic regression analysis when there is complete separation between the two outcome groups

3. Identify and troubleshoot logistic regression analysis when there is collinearity between variables

Wrap-up

- With experience and time, you won't need to rely so heavily on the systematic approaches to model building and numerical problems
  - This is the experience part of the quote!

- 4-minute exit ticket

- **Next class**
  - Assessing model fit

smoke10
educ only

Class 13 Exit Ticket

https://forms.office.com/r/kWwYz6cs5v