Model Specification

A gender analysis-based teaching example

Gabe Murchison and Emily Unger

Background

- In public health, we typically treat sex/gender as a binary variable (male/female)
- However, sex and gender have multiple dimensions, including...
    - Sex assigned at birth
    - Gender identity
    - Gender expression
- Any of these dimensions may each affect health outcomes, either separately or in interaction with one another
- When measured and modeled correctly, these and other dimensions can help us understand the mechanisms relating gender to the outcomes of interest

Background

- The Youth Risk Behavior Surveillance System is a national school-based survey of adolescent health.

  In 2013 and 2015, four large municipalities included a question about **gender expression**:
  - "A person's appearance, style, dress, or the way they walk or talk may affect how people describe them. How do you think other people at school would describe you?"
  - Response options: Very feminine (1); Mostly feminine (2); Somewhat feminine (3); Equally feminine and masculine (4); Somewhat masculine (5); Mostly masculine (6); Very masculine (7)
- Students were also asked "What is your sex?" (Response options: male or female)
- Outcome: **feeling sad or hopeless** almost every day for two weeks (binary yes/no)
- We will specify and contrast a variety of logistic regression models

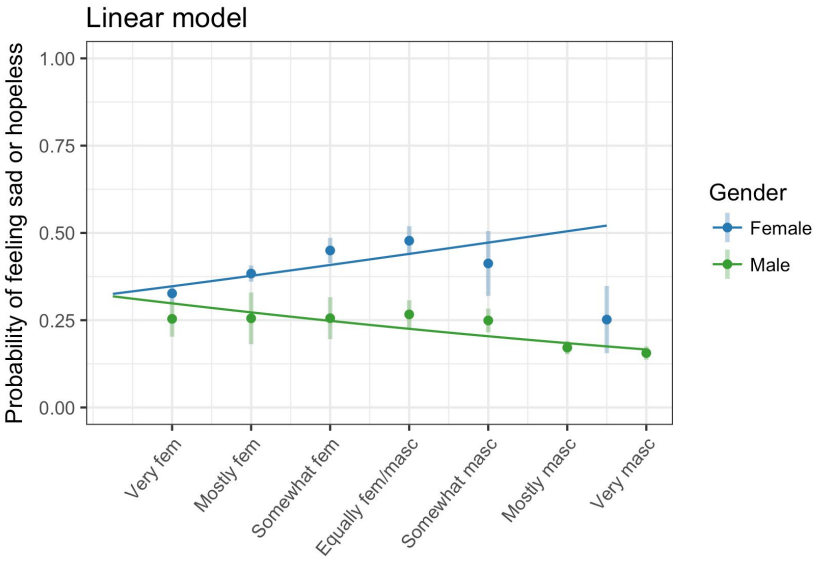
Before Fitting a Model

- What theoretical framework informs your research? How does this shape the substantive questions you are asking?
- What variables should you include in your model? Do you have the data you need to include all the relevant variables? If not, can you get it? If not, what might be invisible in your results?
- What kinds of heterogeneity might exist within or between levels of a given variable?
- For a given variable, what questions do you want to be able to answer?
    - Is it the exposure of interest? If so, interpretability may be a priority.
    - Is it a covariate? If so, eliminating residual confounding may be a priority.
    - Does interpolation matter? Is it meaningful, given the data at hand?

For this data set...

- What might we hypothesize about the relationship between sadness/hopelessness and gender expression? Any competing hypotheses?
- If our hypothesis is true, what would we expect our fitted model to look like?
- Are there any potentially important effect modifiers?

Linear model



Note: although this is a logistic regression model, the fit lines appear linear because the probabilities modeled are not near 0 or 1.

Advantages:

- Easy to interpret
- Allows interpolation

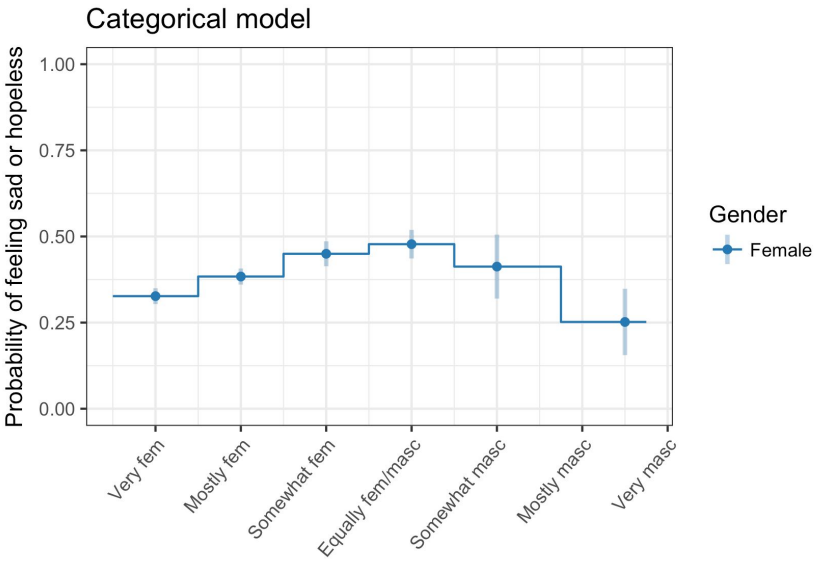
Disadvantages:

- Not flexible
- May hide nonlinear pattern in the data
  - If we modeled these data as linear without visualizing our data, what would we have missed?

$$\ln\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 x_{masc} + \beta_2 x_{female} + \beta_3 x_{masc} x_{female}$$

*Data simulated based on *Gill AM and Frazer MS. 2016. Health Risk Behaviors among Gender Expansive Students: Making the Case for Including a Measure of Gender Expression in Population-Based Surveys. Washington, DC: Advocates for Youth.*

Categorical model



Note: To simplify the model, we are modeling only females.

Advantages:

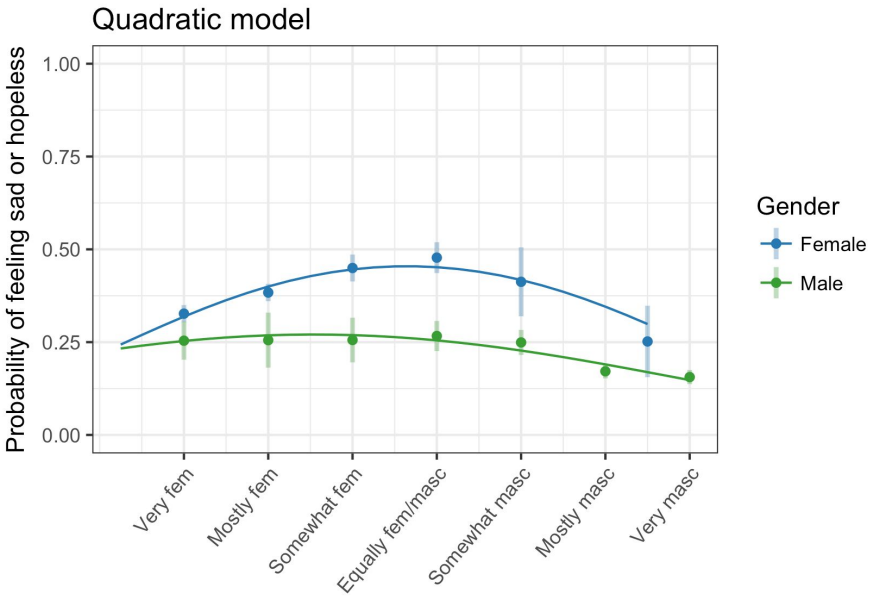
- Doesn't impose any functional form
- Easy to interpret
- Estimates for one category not influenced by those for any other categories

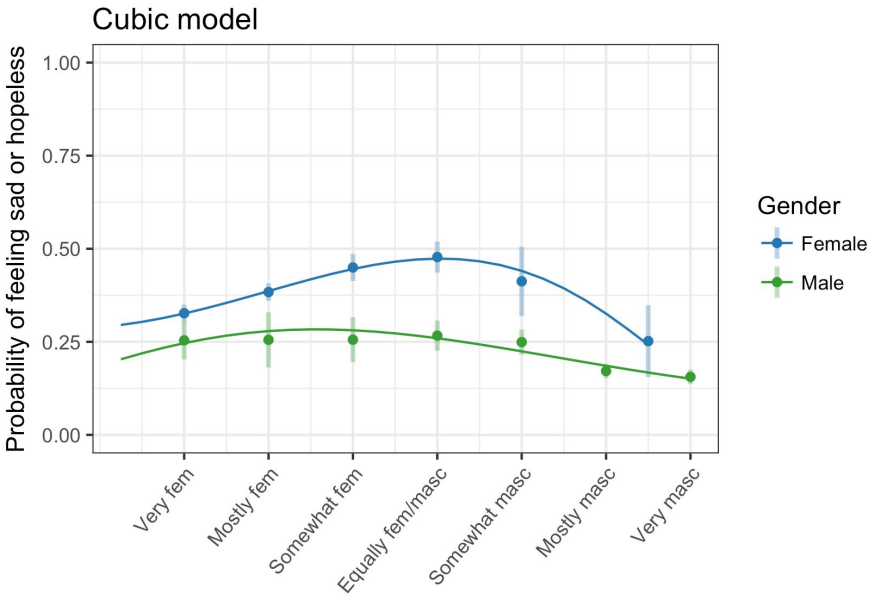
Disadvantages:

- Cannot interpolate
- Comparisons refer to reference category
  - Who should be the reference?
- Consumes more degrees of freedom (less power)
  - Especially if including interaction terms
- Small *n* within categories can cause unstable estimates
- Possibility of heterogeneity within categories
  - How do you choose the categories?

$$\ln\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 x_{masc=2} + \beta_2 x_{masc=3} + \beta_3 x_{masc=4} + \beta_4 x_{masc=5} + \beta_5 x_{masc=6} + \beta_6 x_{masc=7}$$

Polynomials

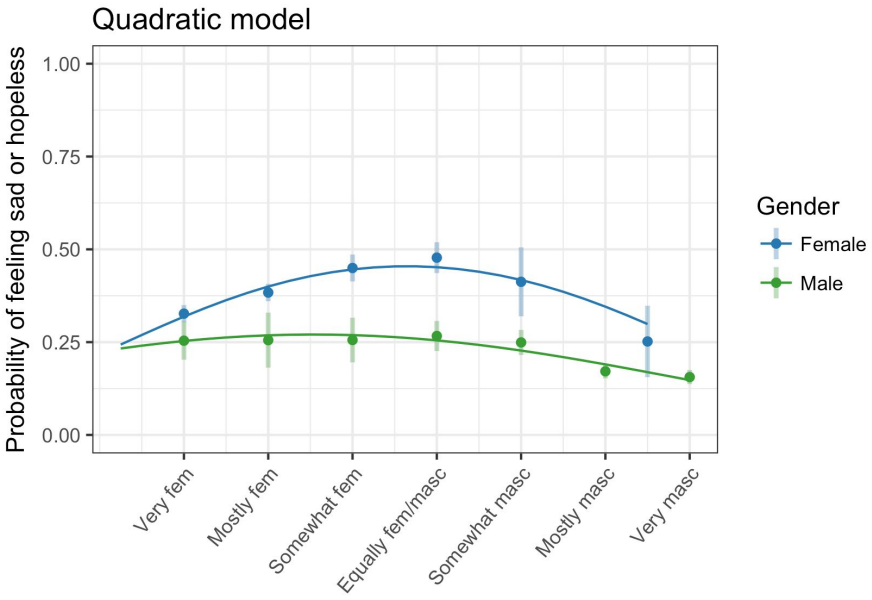




$$\ln\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 x_{masc} + \beta_2 x_{female} + \beta_3 x_{masc}^2 + \beta_4 x_{masc} x_{female} + \beta_5 x_{masc}^2 x_{female}$$

$$\ln\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 x_{masc} + \beta_2 x_{female} + \beta_3 x_{masc}^2 + \beta_4 x_{masc}^3 + \beta_5 x_{masc} x_{female} + \beta_6 x_{masc}^2 x_{female} + \beta_7 x_{masc}^3 x_{female}$$

Polynomials: quadratic model



$$\ln\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 x_{masc} + \beta_2 x_{female} + \beta_3 x_{masc}^2 + \beta_4 x_{masc} x_{female} + \beta_5 x_{masc}^2 x_{female}$$

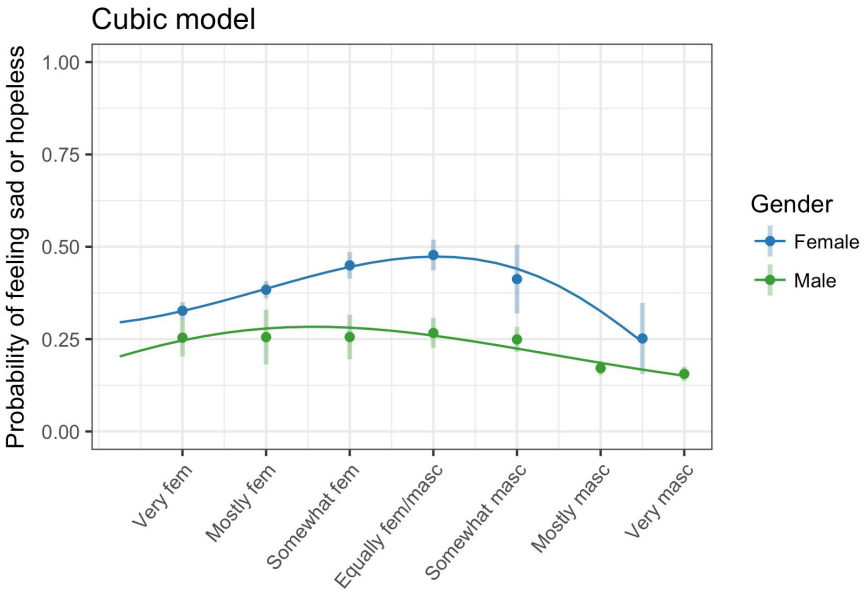
Advantages:

- Relatively simple
- Moderately flexible
- Doesn't consume many degrees of freedom
- Allows interpolation

Disadvantages:

- Somewhat tricky to interpret
- Doesn't always fit data well, especially at extremes
- Estimates at any value of x are influenced by the whole range of x; won't deal well with sudden changes in the x-y relationship (nonlinearities)
- Sensitive to outliers

Polynomials: cubic model



$$\ln\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 x_{masc} + \beta_2 x_{female} + \beta_3 x_{masc}^2 + \beta_4 x_{masc}^3 + \beta_5 x_{masc} x_{female} + \beta_6 x_{masc}^2 x_{female} + \beta_7 x_{masc}^3 x_{female}$$

Advantages and disadvantages:

- Similar to quadratic model
- For polynomial models in general: tradeoff between interpretability and flexibility
- Risk of overfitting
  - Is interpolation meaningful?

Conclusions

- Visualize your data!
  - A linear model would have hidden important findings
- Model choice depends on both the empirical data and the substantive question/theory: what question(s) are you asking, and why?
  - Does interpolation matter?
  - What variables should you include?
    - What would we have missed if our model had only used “sex” and not gender expression?
    - What might still be missing because our data do not distinguish sex assigned at birth and gender?
  - What is the role of a given variable in your study?
    - Is the goal to understand an exposure, or to control confounding?

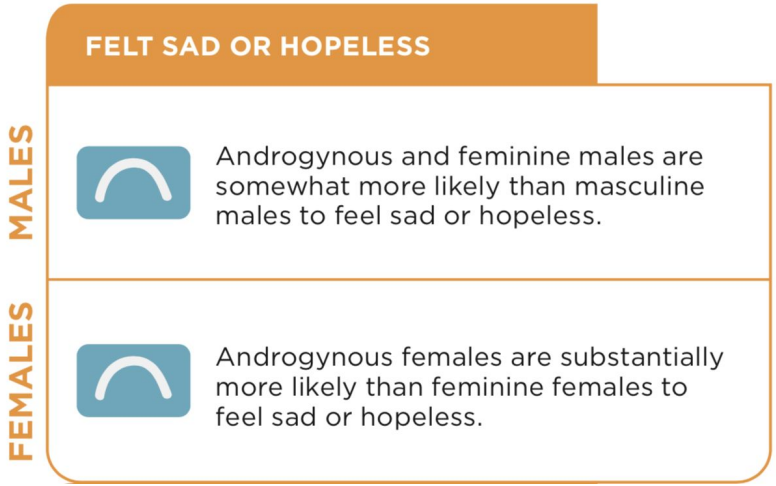


Figure: Gill and Frazer, 2016

- Whatever model you choose, consider how to present your data to facilitate interpretation

References

- For more findings from the YRBSS surveys that included gender expression, and more examples of how to convey complex models in an accessible way, see:
    - Gill AM and Frazer MS. 2016. Health Risk Behaviors among Gender Expansive Students: Making the Case for Including a Measure of Gender Expression in Population-Based Surveys. Washington, DC: Advocates for Youth. Full report available at: http://advocatesforyouth.org/storage/advfy/documents/YRBSS.pdf
- The Youth Risk Behavior Surveillance System: https://www.cdc.gov/healthyyouth/data/yrbs/index.htm

- With thanks to Jarvis Chen and the rest of the PHS 2000A teaching staff!