Exploring the Titanic data

What we've covered so far...

- Interpretation and model fitting (MLE, Fisher scoring, gradient ascent)
- Visualizations and diagnostics (empirical logit plots, quantile residual plots, VIFs, Cook's distance)
- Hypothesis testing (Wald tests)

Data

Data on the RMS *Titanic* disaster. We have data on 891 passengers on the ship, with the following variables:

- `Passenger`: A unique ID number for each passenger.
- `Survived`: An indicator for whether the passenger survived (1) or perished (0) during the disaster.
- `Pclass`: Indicator for the class of the ticket held by this passengers; 1 = 1st class, 2 = 2nd class, 3 = 3rd class.
- `Sex`: Binary Indicator for the biological sex of the passenger.
- `Age`: Age of the passenger in years; Age is fractional if the passenger was less than 1 year old.
- `Fare`: How much the ticket cost in US dollars.
- `+` others

Research question

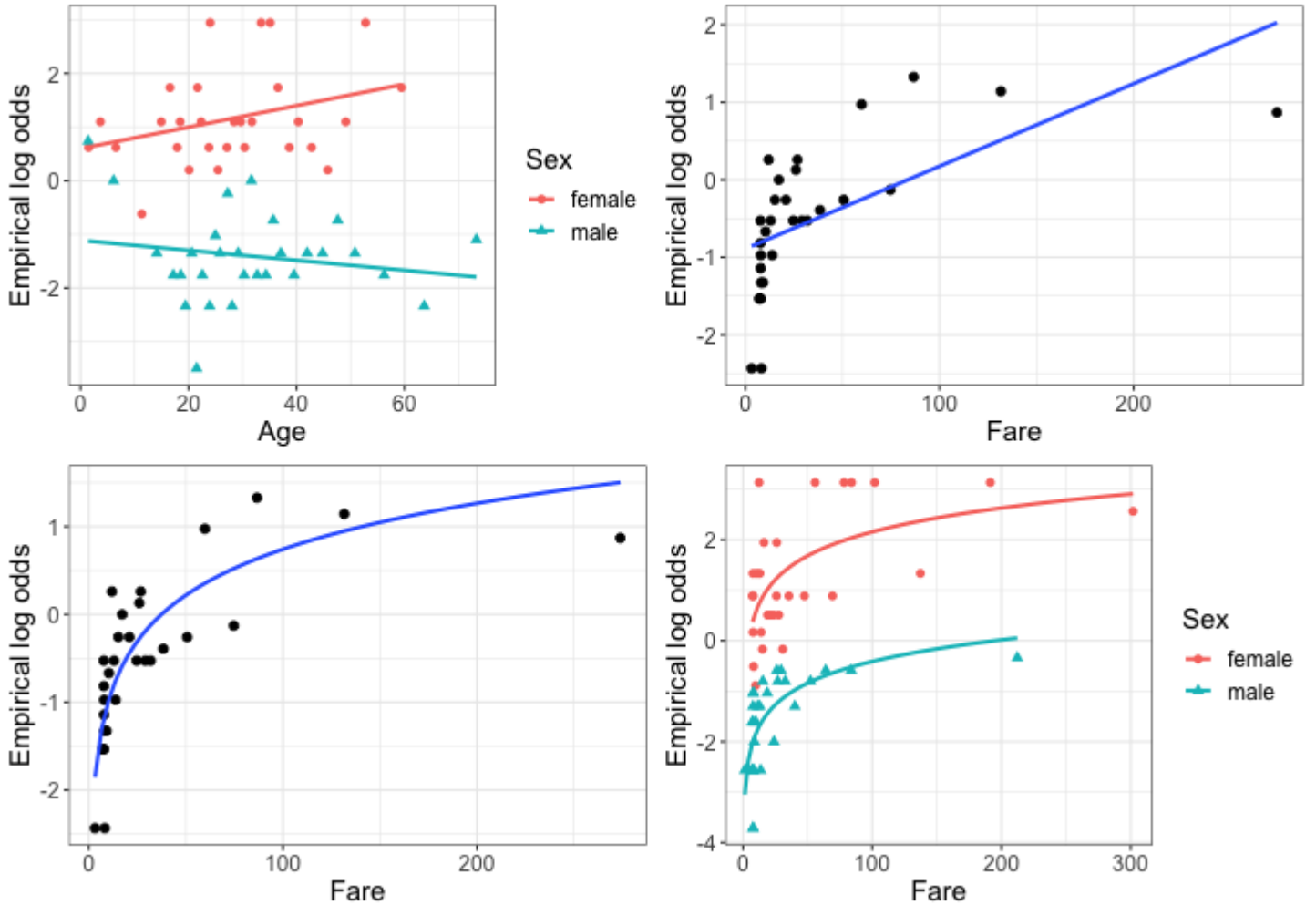
Is there a relationship between passenger age and their probability of survival, after accounting for sex, passenger class, and the cost of their ticket?

> What steps should I take to investigate this question with logistic regression?

Class activity, Part I (EDA)

https://sta712-f22.github.io/class_activities/ca_lecture_12.html

Class activity



Class activity

> Based on your EDA, what model would you fit to address the research question?

Class activity, Part II (Diagnostics)

https://sta712-f22.github.io/class_activities/ca_lecture_12.html

Class activity, Part III (Hypothesis testing)

https://sta712-f22.github.io/class_activities/ca_lecture_12.html