Lecture 28

Recap: multinomial regression model

Data $(X_1, Y_1), \ldots, (X_n, Y_n)$

$$Y_i \sim \text{Categorical}(\pi_{i1}, \ldots, \pi_{iJ})$$

$$\mu_i = (\pi_{i1}, \ldots, \pi_{i,J-1})^T \in \mathbb{R}^{J-1}$$

$$g(\mu_i) = \begin{pmatrix} \log\left(\frac{\pi_{i1}}{1 - \sum_{j=1}^{J} \pi_{ij}}\right) \\ \vdots \\ \log\left(\frac{\pi_{i,J-1}}{1 - \sum_{j=1}^{J-1} \pi_{ij}}\right) \end{pmatrix} = \begin{pmatrix} \beta_1^T X_i \\ \vdots \\ \beta_{J-1}^T X_i \end{pmatrix}$$

$$= X_i^* \beta$$

$$\begin{bmatrix} X_i^T & & \\ & X_i^T & \\ & \cdots & X_i^T \end{bmatrix} \qquad \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{J-1} \end{bmatrix}$$

Motivating example: earthquake data

We have data from the 2015 Gorkha earthquake in Nepal. After the earthquake, a large scale survey was conducted to determine the amount of damage the earthquake caused for homes, businesses and other structures. Variables include:

$$\text{Damage}_i \sim \text{Categorical}\left(\pi_{i(\text{none})}, \pi_{i(\text{mod})}, \pi_{i(\text{severe})}\right)$$

- `Damage`: the amount of damage suffered by the building (none, moderate, severe)
- `age`: the age of the building (in years)
- `condition`: a de-identified variable recording the condition of the land surrounding the building

$$\log\left(\frac{\pi_{i(\text{mod})}}{\pi_{i(\text{none})}}\right) = \beta_{(\text{mod})}^T x_i$$

$$\log\left(\frac{\pi_{i(\text{severe})}}{\pi_{i(\text{none})}}\right) = \beta_{(\text{severe})}^T x_i$$

Exploratory data analysis

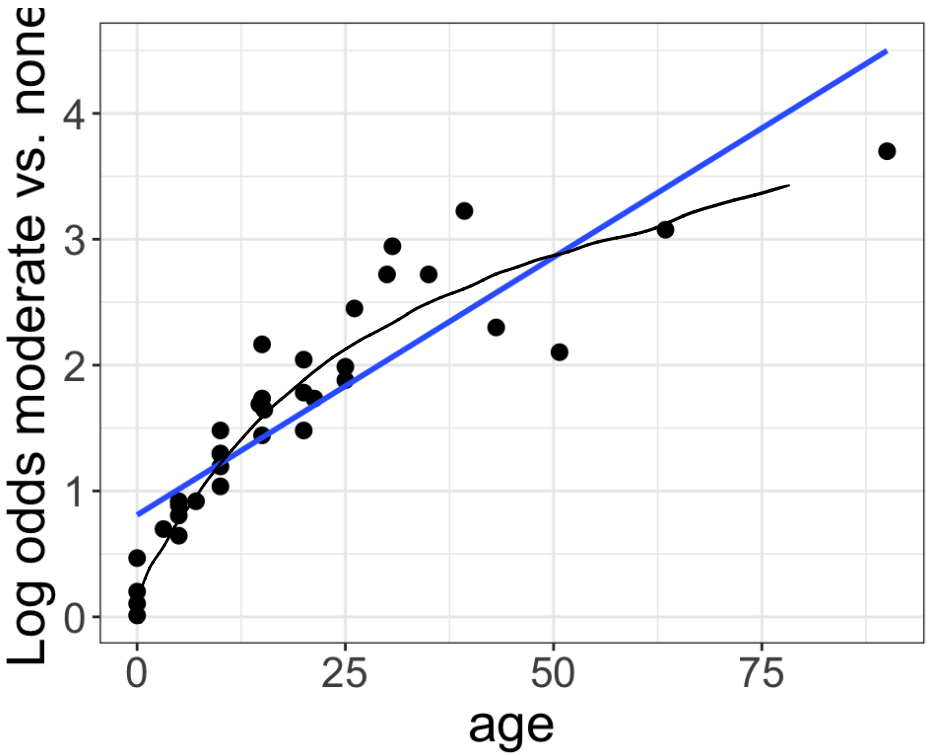
We want to model damage using age and land surface condition. What kind of EDA could I do?

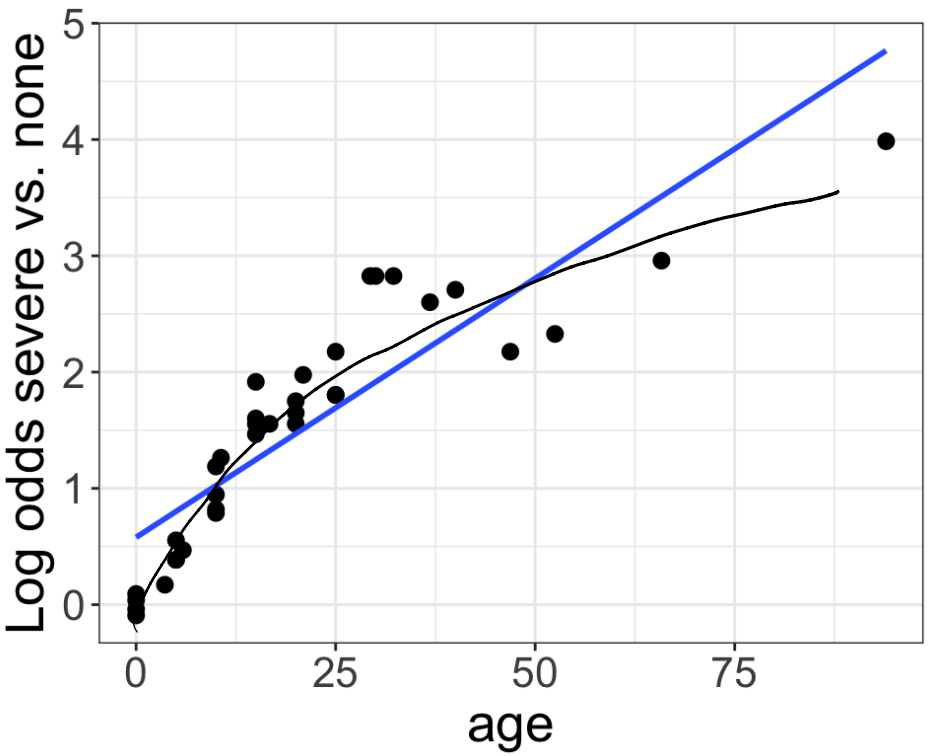
Empirical logit plots!

Compare Moderate vs. None

Severe vs. None

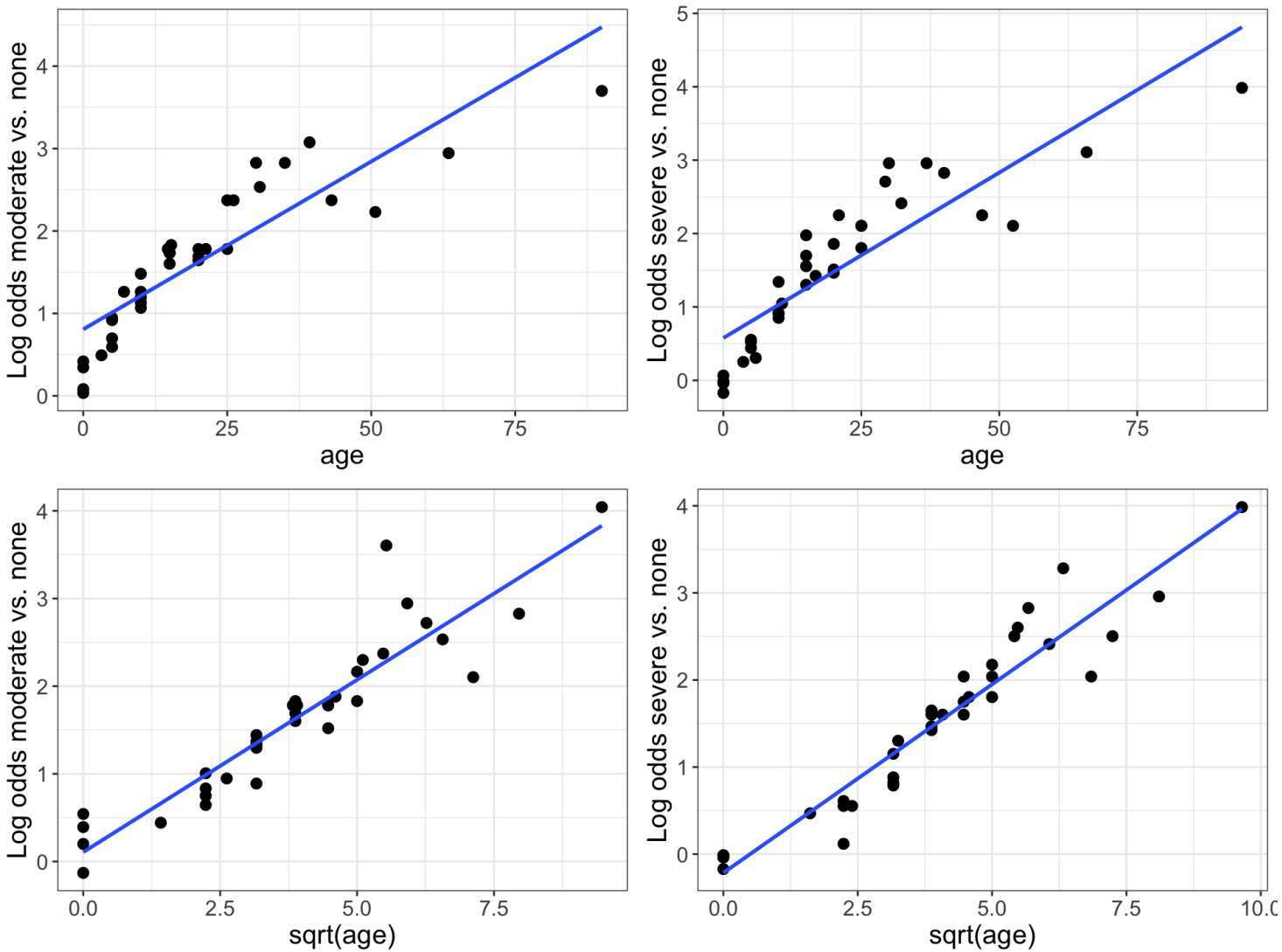
Empirical logit plots





Try a transformation? Square root?

Trying a transformation



Fitting the model in R

fits multinomial regression

```
1 library(nnet)
2 m1 <- multinom(Damage ~ sqrt(age) +
3                   condition,
4                 data = earthquake)
```

```
1 summary(m1)
```

```
...
Coefficients:
         (Intercept) sqrt(age)  conditiono conditiont
moderate   0.6581163 0.3747641 -0.45376940 -0.5803708
severe     0.1881145 0.4251732  0.04706934 -0.4623774

Std. Errors:
         (Intercept)  sqrt(age) conditiono conditiont
moderate   0.1208913 0.01684468  0.2305975  0.1155475
severe     0.1243799 0.01725782  0.2292533  0.1180182
...
```

coefficients for logits

Standard errors for coefficients

$$\log\left(\frac{\hat{\pi}_{i(\text{mod})}}{\hat{\pi}_{i(\text{None})}}\right) = 0.658 + 0.375\sqrt{\text{Age}_i} - 0.454\,O_i - 0.580\,T_i$$

A one-unit increase in $\sqrt{\text{Age}_i}$ is associated w/ an increase in the odds of moderate vs. no damage by a factor of $\exp\{0.375\} = 1.45$, holding surface condition fixed

Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_28.html

Class activity

1) odds: $\exp\{0.658 + 0.375\sqrt{25} - 0.454\}$

≈ 8

(moderate is 8 times as likely as no damage)

2) $\hat{\pi}_{i(\text{moderate})}$

In general: $\hat{\pi}_{ij} = \dfrac{\text{odds}(j \text{ vs. } J)}{1 + \sum_{h=1}^{J-1} \text{odds}(h \text{ vs. } J)}$

$\dfrac{p_M}{p_n} = 7.996$

by similar logic:

$\dfrac{p_S}{p_n} = 10.591$

$\dfrac{p_n}{p_n} = 1$

implies $7.996 p_n = p_M$

$10.591 p_n = p_S$

$p_M + p_n + p_S = 1$

$p_n(7.996 + 10.591 + 1) = 1 \Rightarrow p_n = .051$

$7.996 p_n = .408 = p_M$ ☺

Fisher scoring for multinomial regression