Lecture 27

Motivating example: earthquake data

Data from the 2015 Gorkha earthquake in Nepal. Variables include:

- `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

Research goal: Build a model to predict Damage

Damage is a categorical variable, w/ > 2 levels

The categorical distribution

Damage has levels None, Moderate, Severe
(ignore ordering for now)

$$\text{Damage}_i \sim \text{Categorical}(\pi_{i(\text{None})}, \pi_{i(\text{Moderate})}, \pi_{i(\text{Severe})})$$

$$\pi_{i(\text{None})} = P(\text{Damage}_i = \text{None}), \text{etc.}$$

In general, $Y \sim \text{Categorical}(\pi_1, \dots, \pi_J)$

when $Y \in \{1, \dots, J\}$ w/ probabilities $\pi_j = P(Y = j)$

(requires: $0 \leq \pi_j \leq 1 \quad \forall j$

$\sum_{j=1}^{J} \pi_j = 1$)

Write Categorical like EDM: $Y \sim \text{Categorical}(\pi_1, \ldots, \pi_J)$

$$f_Y(y; \pi_1, \ldots, \pi_J) = \begin{cases} \pi_1 & y=1 \\ \pi_2 & y=2 \\ \vdots & \\ \pi_J & y=J \end{cases} = \prod_{j=1}^{J} \pi_j^{\mathbb{1}\{y=j\}}$$

Let $y_j^* = \begin{cases} 1 & y=j \\ 0 & y \neq j \end{cases} = \mathbb{1}\{y=j\}$

$$f(y; \pi_1, \ldots, \pi_J) = \prod_{j=1}^{J} \pi_j^{y_j^*} = \left(\prod_{j=1}^{J-1} \pi_j^{y_j^*}\right)\left(1 - \sum_{j=1}^{J-1} \pi_j\right)^{y_J^*}$$

$$= \left(\prod_{j=1}^{J-1} \pi_j^{y_j^*}\right)\left(1 - \sum_{j=1}^{J-1} \pi_j\right)^{1 - \sum_{j=1}^{J-1} y_j^*}$$

$$\pi_J = 1 - \sum_{j=1}^{J-1} \pi_j$$

$$\sum_{j=1}^{J} y_j^* = 1$$

Bernoulli: $p^y(1-p)^{1-y}$

$$f(y; \pi_1, \dots, \pi_{J-1}) = \left(\prod_{j=1}^{J-1} \pi_j^{y_j^*}\right)\left(1 - \sum_{j=1}^{J-1} \pi_j\right)^{1-\sum_{j=1}^{J-1} y_j^*}$$

$$= \exp\left\{ \sum_{j=1}^{J-1} y_j^* \log \pi_j + \left(1 - \sum_{j=1}^{J-1} y_j^*\right) \log\left(1 - \sum_{j=1}^{J-1} \pi_j\right) \right\}$$

$$= \exp\left\{ \underbrace{\sum_{j=1}^{J-1} y_j^* \log\left(\frac{\pi_j}{1 - \sum_{k=1}^{J-1} \pi_k}\right)}_{y^{*T}\theta} + \underbrace{\log\left(1 - \sum_{j=1}^{J-1} \pi_j\right)}_{-k(\theta)} \right\}$$

$\phi = 1 \in \mathbb{R}$ $\quad a(y^*, \phi) = 1 \in \mathbb{R}$

$$\theta \in \mathbb{R}^{J-1} = \left(\log\left(\frac{\pi_1}{1 - \sum_{k=1}^{J-1} \pi_k}\right), \dots, \log\left(\frac{\pi_{J-1}}{1 - \sum_{k=1}^{J-1} \pi_k}\right) \right)^T$$

$k : \mathbb{R}^{J-1} \to \mathbb{R}$

$$y^* \in \mathbb{R}^{J-1} = (y_1^*, \dots, y_{J-1}^*)^T$$

Multivariate EDM :

$$f(y^* ; \theta, \phi) = a(y^*, \phi) \exp\left\{ \frac{y^{*T}\theta - \kappa(\theta)}{\phi} \right\}$$

y^*: vector θ: vector ϕ: scalar > 0

$\kappa(\theta) \in \mathbb{R}$, $a(y^*, \phi) \in \mathbb{R}$

Multivariate GLM

Suppose we observe data $(X_1, Y_1), \ldots, (X_n, Y_n)$

$X_i \in \mathbb{R}^p \qquad Y_i \sim \text{Categorical}(\pi_{i1}, \ldots, \pi_{iJ})$

$Y^*_{ij} = \mathbb{1}\{Y_i = j\} \Rightarrow Y^*_i = (Y^*_{i1}, \ldots, Y^*_{i,J-1})^T \in \mathbb{R}^{J-1}$

$$\mathbb{E}[Y^*_i] = \mu_i = \begin{bmatrix} \pi_{i1} \\ \vdots \\ \pi_{i,J-1} \end{bmatrix} \in \mathbb{R}^{J-1}$$

$\beta_1 = (\beta_{10}, \ldots, \beta_{1,p-1})^T$

$\beta_2 = (\beta_{20}, \ldots, \beta_{2,p-1})^T$

$\vdots$

$$g_1(\mu_i) = \beta_1^T X_i$$
$$g_2(\mu_i) = \beta_2^T X_i$$
$$\vdots$$
$$g_{J-1}(\mu_i) = \beta_{J-1}^T X_i$$

$g_j : \mathbb{R}^{J-1} \to \mathbb{R}$

$$g(\mu_i) = \begin{pmatrix} g_1(\mu_i) \\ \vdots \\ g_{J-1}(\mu_i) \end{pmatrix} \in \mathbb{R}^{J-1}$$

$g : \mathbb{R}^{J-1} \to \mathbb{R}^{J-1}$

$$g(\mu_i) = \begin{pmatrix} g_1(\mu_i) \\ \vdots \\ g_{J-1}(\mu_i) \end{pmatrix} \in \mathbb{R}^{J-1}$$

$$= \begin{bmatrix} \beta_1^T x_i \\ \beta_2^T x_i \\ \vdots \\ \beta_{J-1}^T x_i \end{bmatrix} = \underbrace{\begin{bmatrix} x_i^T & & & 0 \\ & x_i^T & & \\ & & \ddots & \\ 0 & & & x_i^T \end{bmatrix}}_{X_i^*} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{J-1} \end{bmatrix}}_{\beta}$$

$$g(\mu_i) = X_i^* \beta$$

Multinomial regression model

$$Y_i \sim \text{Categorical}(\pi_1, \dots, \pi_J)$$

$$g(\mu_i) = \theta_i = \begin{pmatrix} \log\left(\frac{\pi_1}{1 - \sum_{j=1}^{J-1} \pi_j}\right) \\ \vdots \\ \log\left(\frac{\pi_{J-1}}{1 - \sum_{j=1}^{J-1} \pi_j}\right) \end{pmatrix}$$

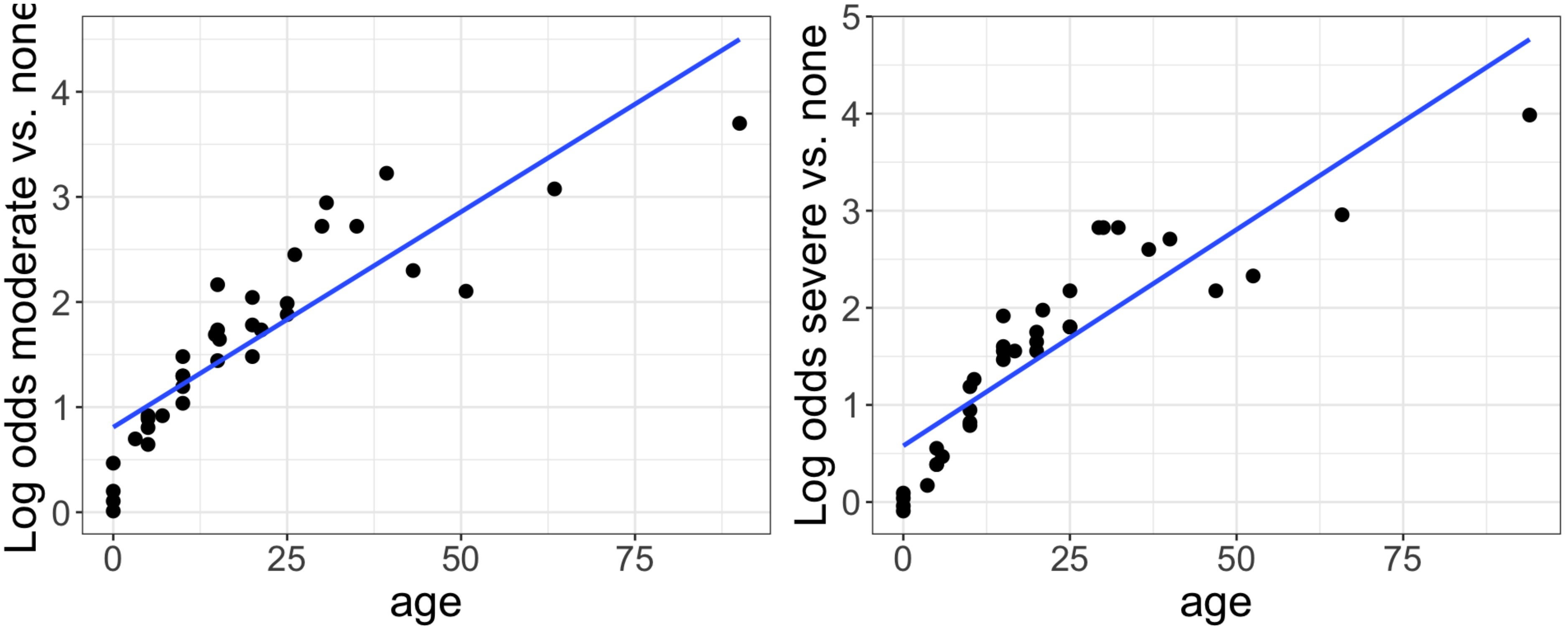
$$= \begin{pmatrix} \log\left(\frac{\pi_1}{\pi_J}\right) \\ \vdots \\ \log\left(\frac{\pi_{J-1}}{\pi_J}\right) \end{pmatrix} = \begin{pmatrix} \beta_1^T X_i \\ \vdots \\ \beta_{J-1}^T X_i \end{pmatrix}$$

baseline-category logits

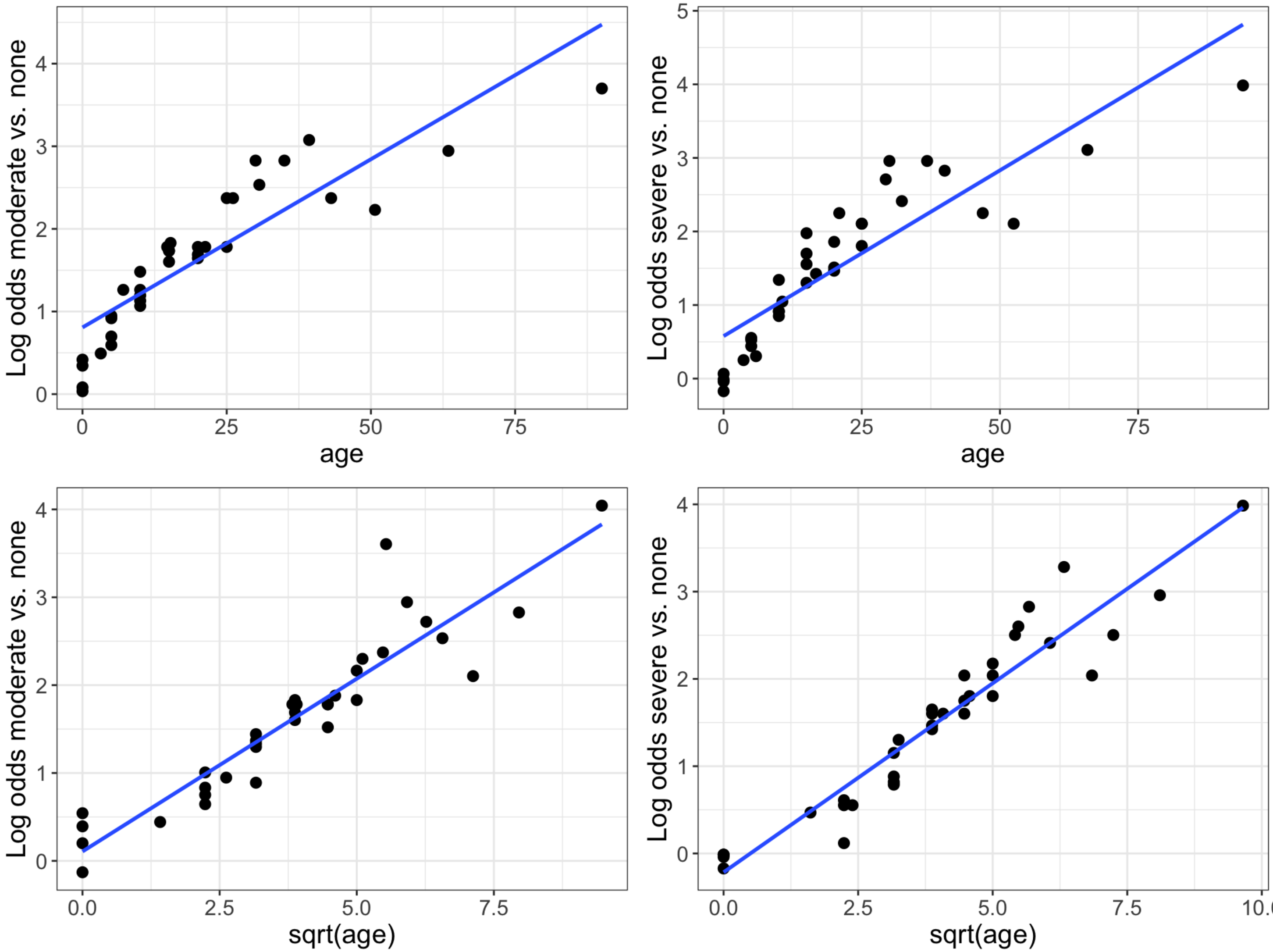
Exploratory data analysis

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

Empirical logit plots



Trying a transformation



Fitting the model in R

```r
library(nnet)
m1 <- multinom(Damage ~ sqrt(age) +
                  condition,
               data = earthquake)
```

```r
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
...
```