Lecture 13

Data

2015 Family Income and Expenditure Survey (FIES) on households in the Phillipines. Variables include

- `age`: age of the head of household
- `numLT5`: number in the household under 5 years old
- `total`: total number of people other than head of household
- `roof`: type of roof (stronger material can sometimes be used as a proxy for greater wealth)
- `location`: where the house is located (Central Luzon, Davao Region, Ilocos Region, Metro Manila, or Visayas)

Poisson regression model

Y_i = number of people in household other than head

$$Y_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Age}_i$$

Age of head of household

Model assumptions

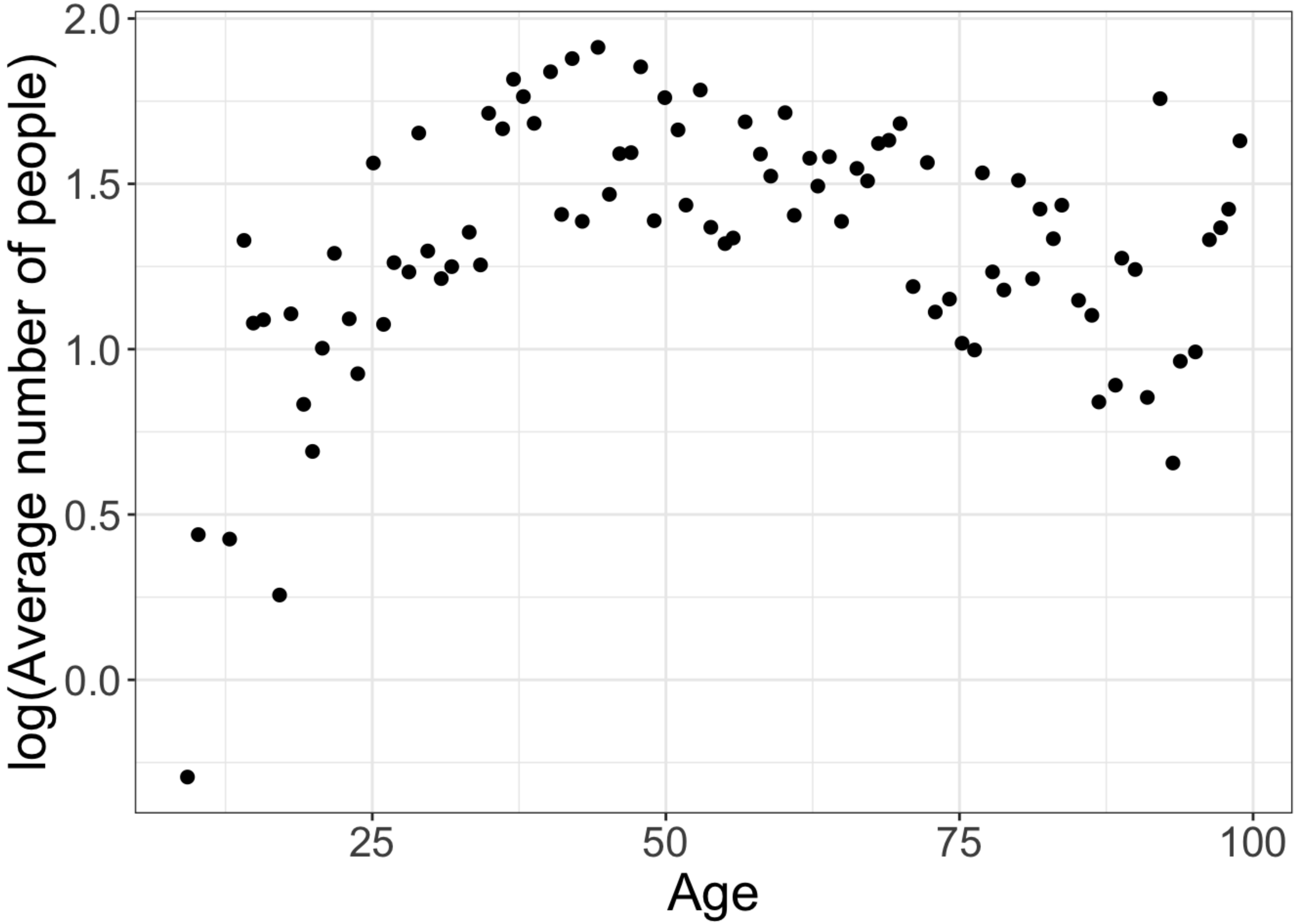
Y_i = number of people in household other than head

$$Y_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \, \text{Age}_i$$

- **Shape:** The shape of the regression model is correct
- **Independence:** The observations are independent
- **Poisson distribution:** A Poisson distribution is a good choice for Y_i

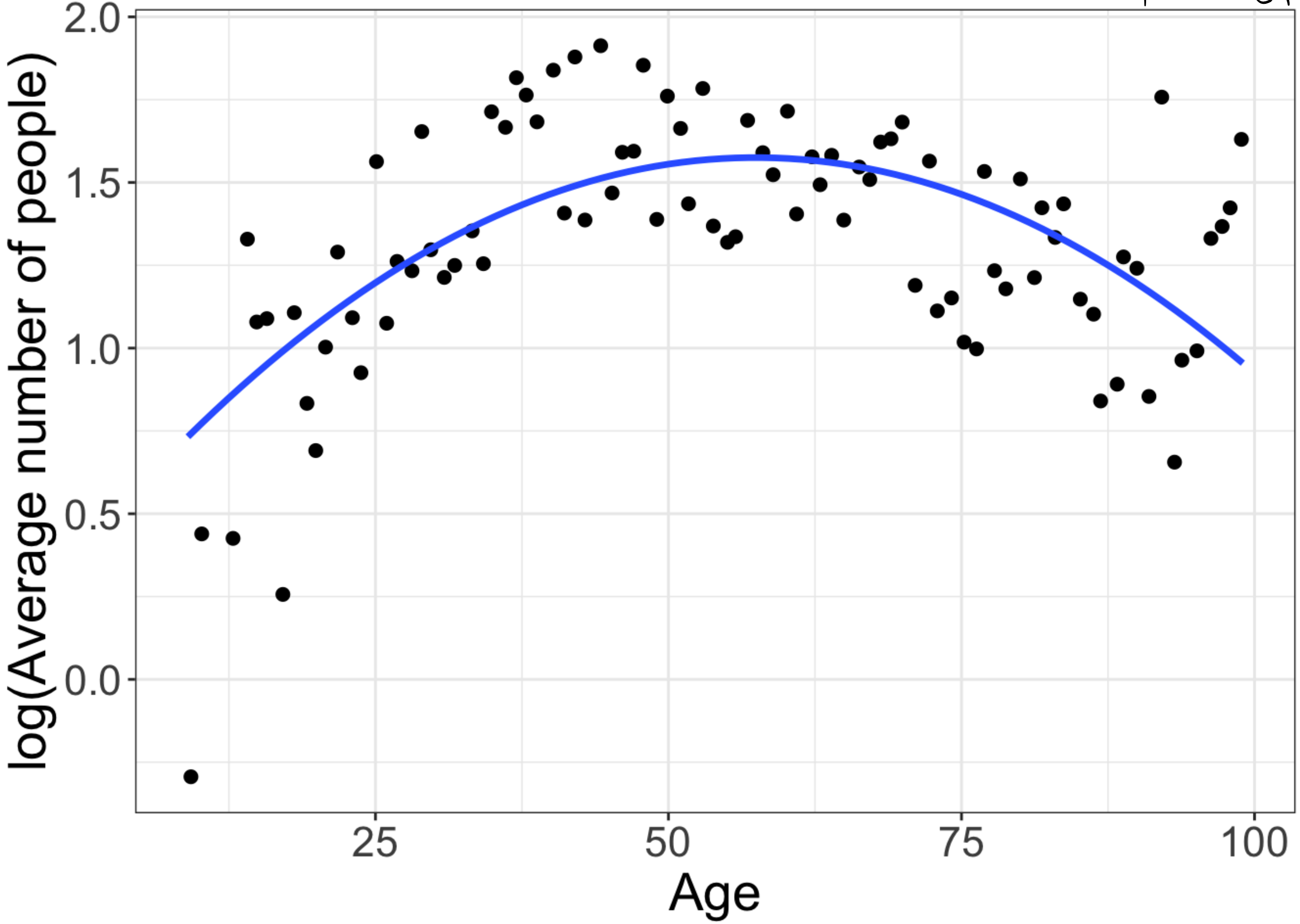
Shape: log empirical means plot



- bin explanatory variables into n_{bins} groups
- calculate $\log(\bar{Y})$ in each bin
- Plots empirical log means against bin centers

Shape: log empirical means plot

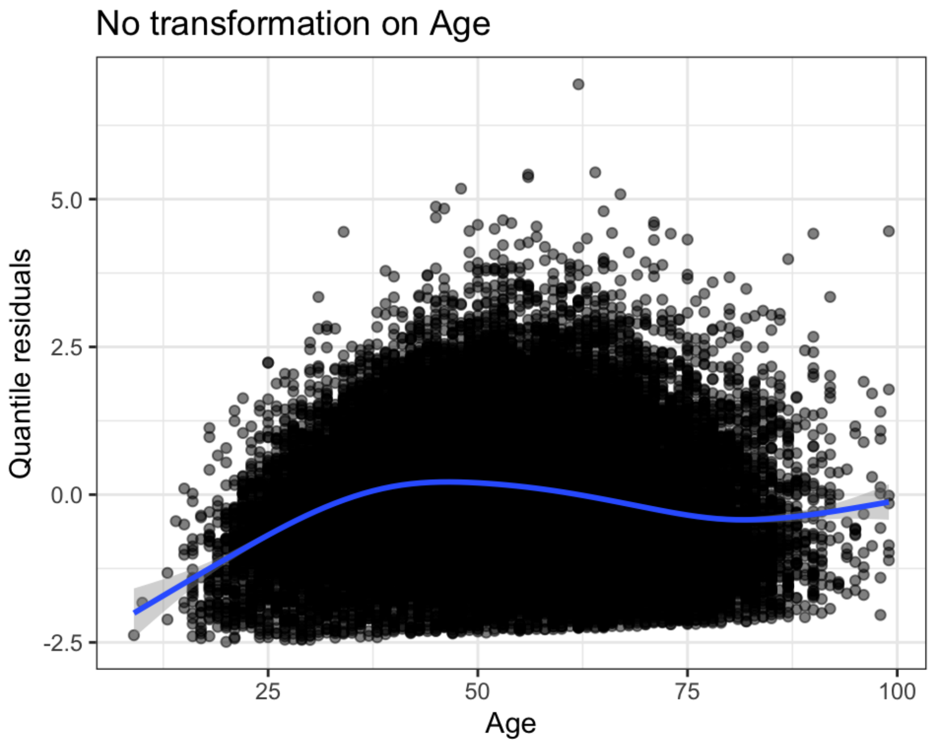
← do in EDA to look for transformations before fitting model

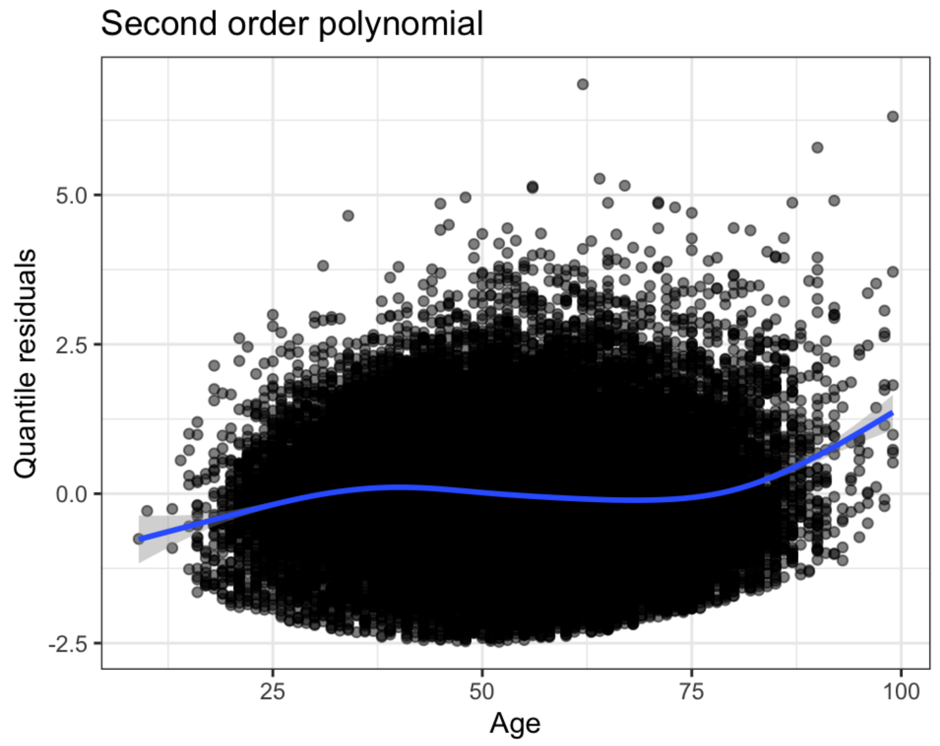


Maybe : $\log(\lambda_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2$

Shape: quantile residual plot

```r
m1 <- glm(total ~ age,
          data = fies, family = poisson)
m2 <- glm(total ~ poly(age, 2),
          data = fies, family = poisson)
```



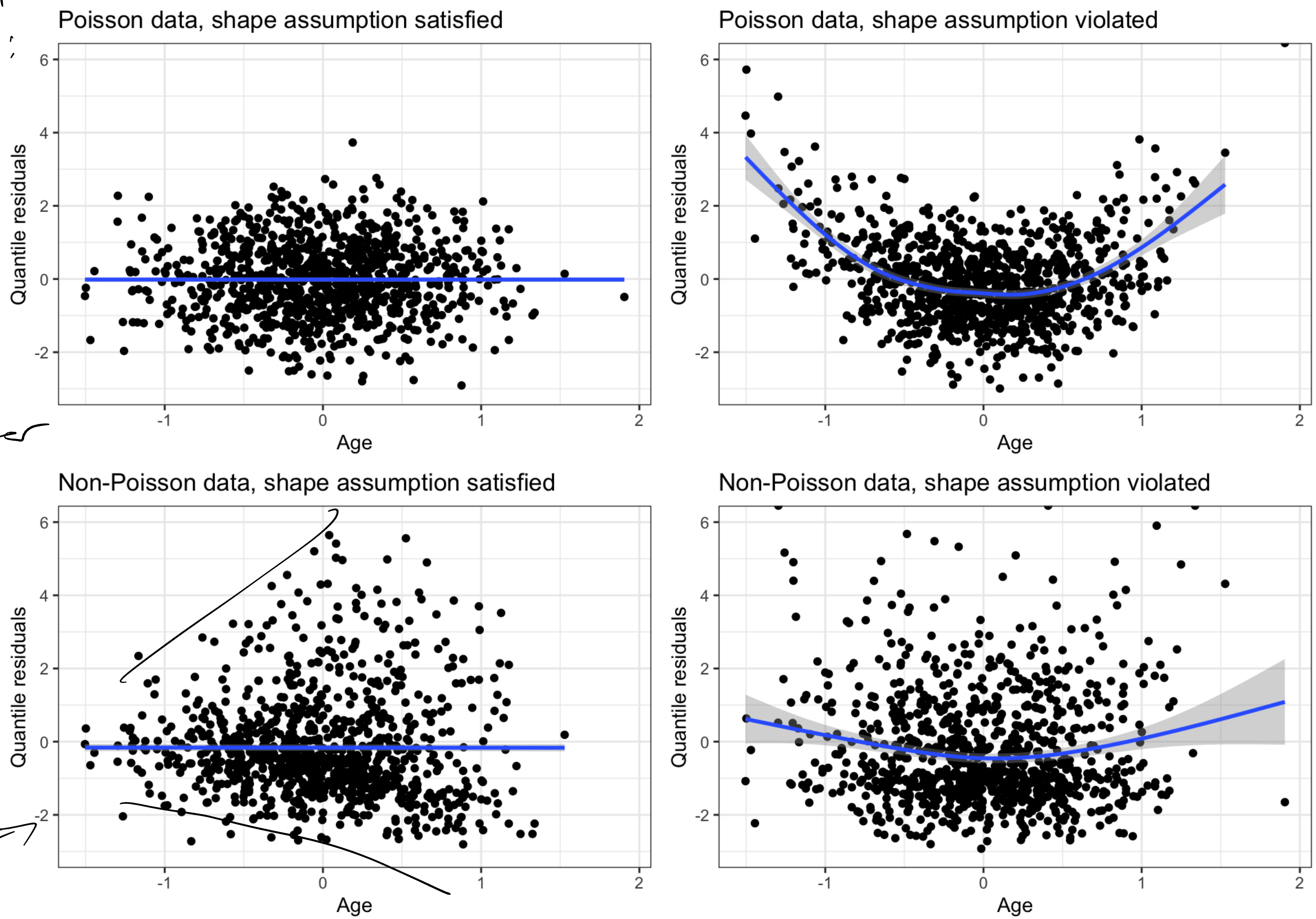


Too much variability and/or nonconstant variance: wrong distribution

Class activity from last time

If distribution + shape is correct:

$r_i \sim N(0,1)$



Pattern suggests violation of shape assumption

· random scatter around 0

· constant variance

· most residuals $\in [-2, 2]$

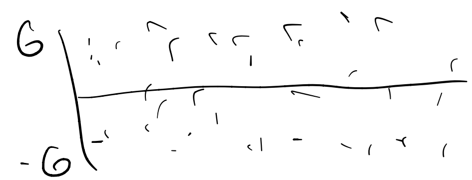
· No pattern

· Non-constant variance

· Too much variability

$Y_i \sim NB$

$\log(\mu_i) = \beta_0 + \beta_1 X_i$



Using quantile residual plots

We can use the quantile residual plot to assess the shape and distribution assumptions:

- Changes in variance indicate potential violations of the distribution assumption
- Patterns indicate potential violations of the shape assumption

A goodness-of-fit test

H_0: model is a good fit to the data

H_A: model is <u>not</u> a good fit to the data

Idea: · find a test statistic with a known distribution when model is correct

· Compare observed test statistic to that distribution

Test statistic: $\dfrac{D(Y, \hat{\mu})}{\phi} \approx \chi^2_{n-p}$ ← # parameters in model

<u>if</u> H_0 is true

Motivation: if $Y \sim N(\mu, \sigma^2)$

$d(y, \mu) = (y - \mu)^2$

$\Rightarrow \dfrac{D(Y, \hat{\mu})}{\phi} = \dfrac{\sum_i (Y_i - \hat{\mu}_i)^2}{\sigma^2} \sim \chi^2_{n-p}$

A goodness-of-fit test

Requirements: saddlepoint approximation

$$f(y; \mu, \phi) = b(y, \phi) \exp\left\{ \frac{-d(y,\mu)}{2\phi} \right\}$$

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\}$$

For many EDMs: $b(y, \phi) \approx \frac{1}{\sqrt{2\pi\phi V(y)}}$

(For Normal: approximation is exact)

For Poisson: saddlepoint approximation is good when $\min y_i \geq 3$

A goodness-of-fit test

```r
m1 <- glm(total ~ age,
          data = fies, family = poisson)

pchisq(m1$deviance, m1$df.residual, lower.tail=F)
```

```
[1] 1.062105e-42
```

```r
m2 <- glm(total ~ poly(age, 2),
          data = fies, family = poisson)

pchisq(m2$deviance, m2$df.residual, lower.tail=F)
```

```
[1] 7.043171e-12
```

Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_13.html