Poisson Regression

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

$$Y_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 Age_i$$

Model assumptions

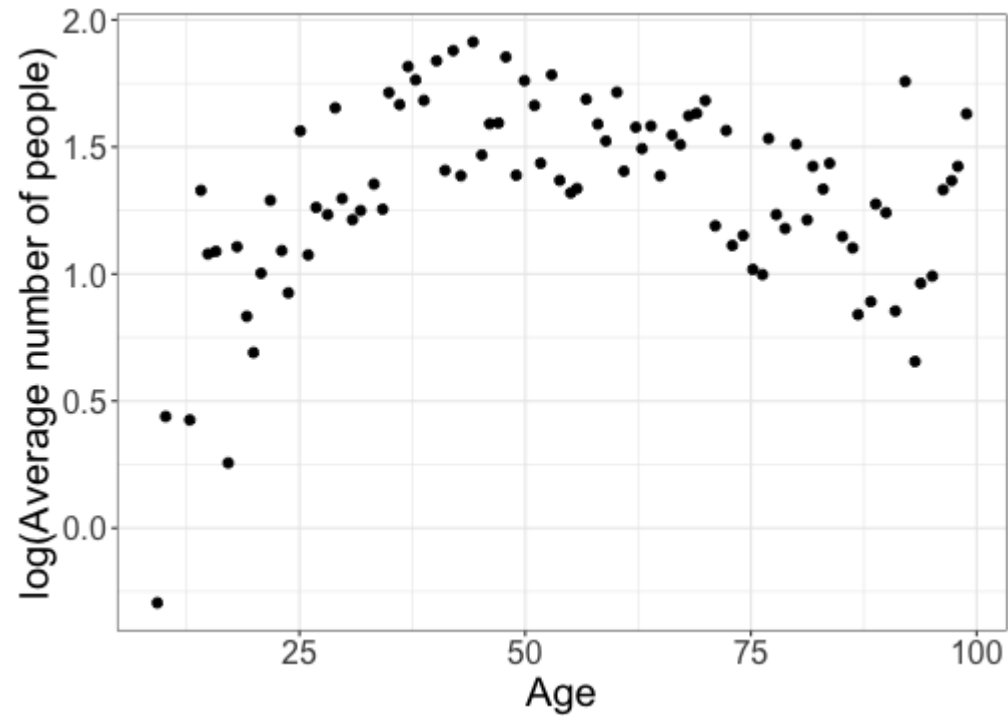$Y_i = $ number of people in household other than head

$$Y_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 Age_i$$

Assumptions:
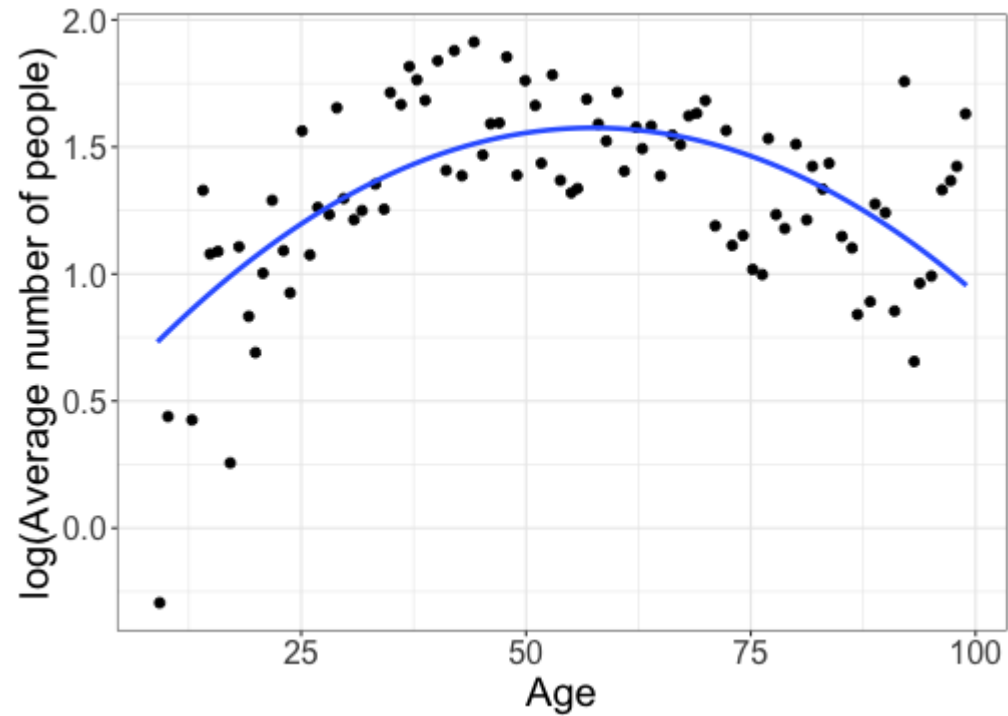
✚ **Shape:** The shape of the regression model is correct

✚ **Independence:** The observations are independent

✚ **Poisson distribution:** A Poisson distribution is a good choice for Y_i

How could I assess these assumptions?
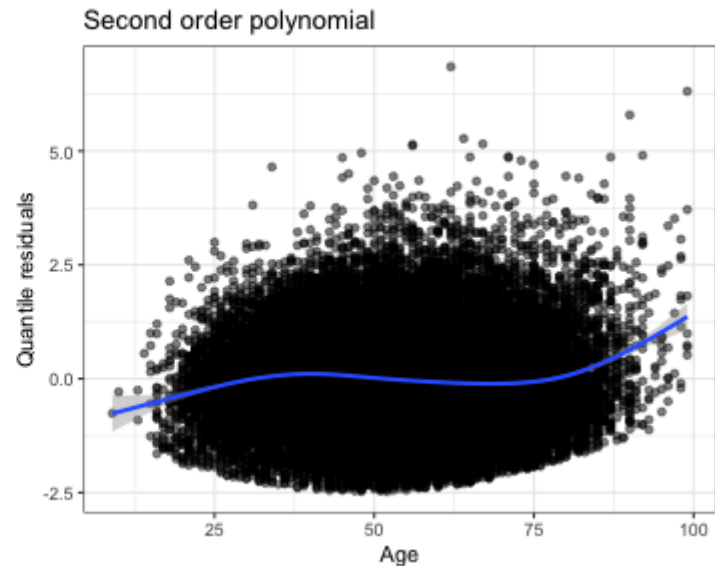
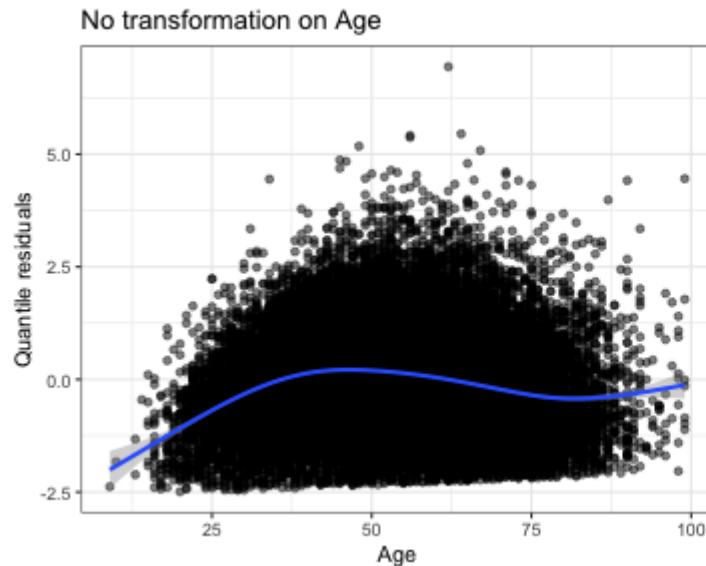# Shape: log empirical means plot

Shape: log empirical means plot

Shape: quantile residual plot

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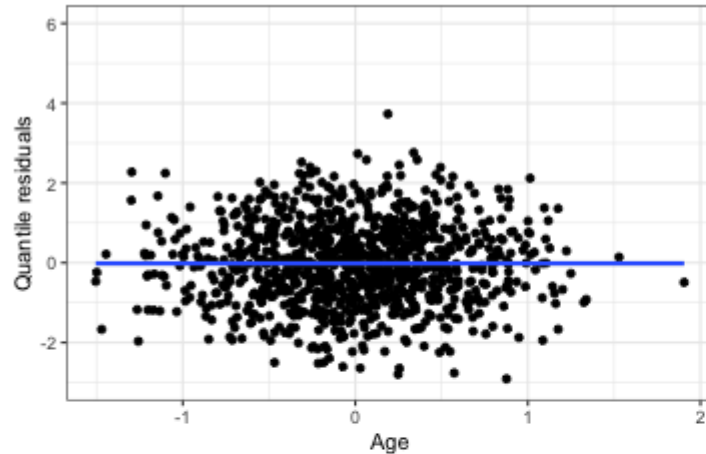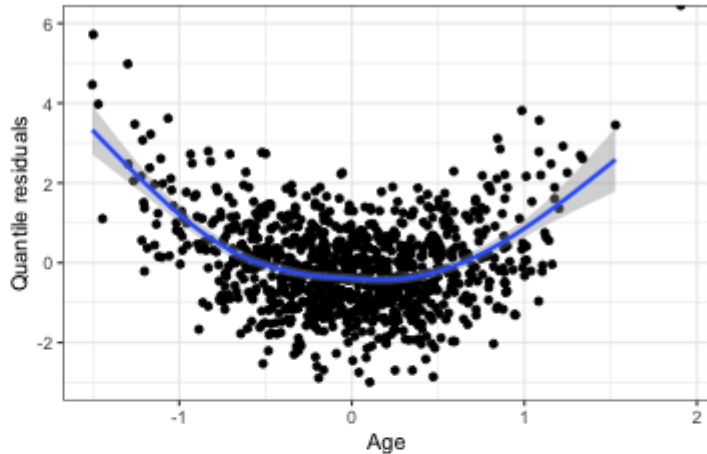
Class activity

https://sta214-s23.github.io/class_activities/ca_lecture_19.html
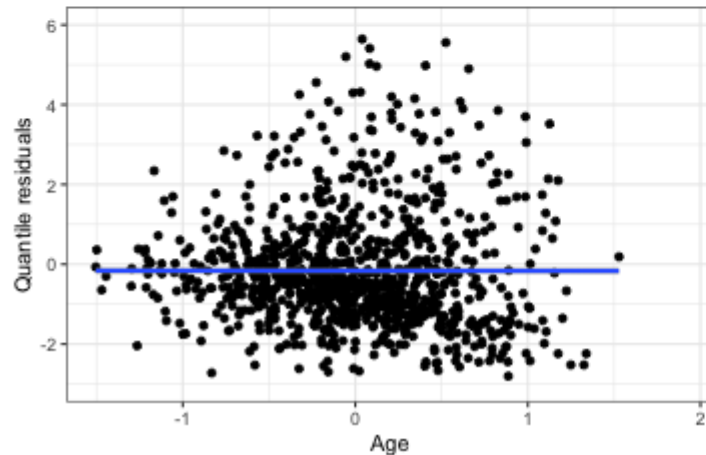
Class activity
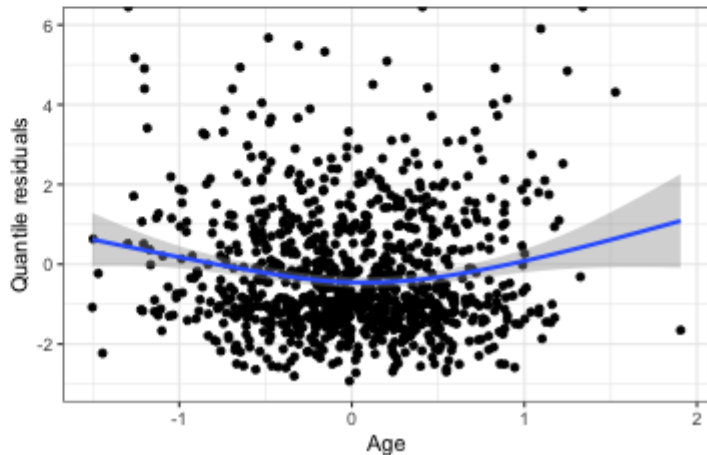
Using quantile residual plots

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

Another dataset

A concerned parent asks us to investigate crime rates on college campuses. We have access to data on 81 different colleges and universities in the US, including the following variables:

➕ `type`: college (C) or university (U)

➕ `nv`: the number of crimes for that institution in the given year

➕ `enroll1000`: the number of enrolled students, in thousands

➕ `region`: region of the US C = Central, MW = Midwest, NE = Northeast, SE = Southeast, SW = Southwest, and W = West)

Question

We want to know whether there are regional differences in the number of crimes on college campuses.

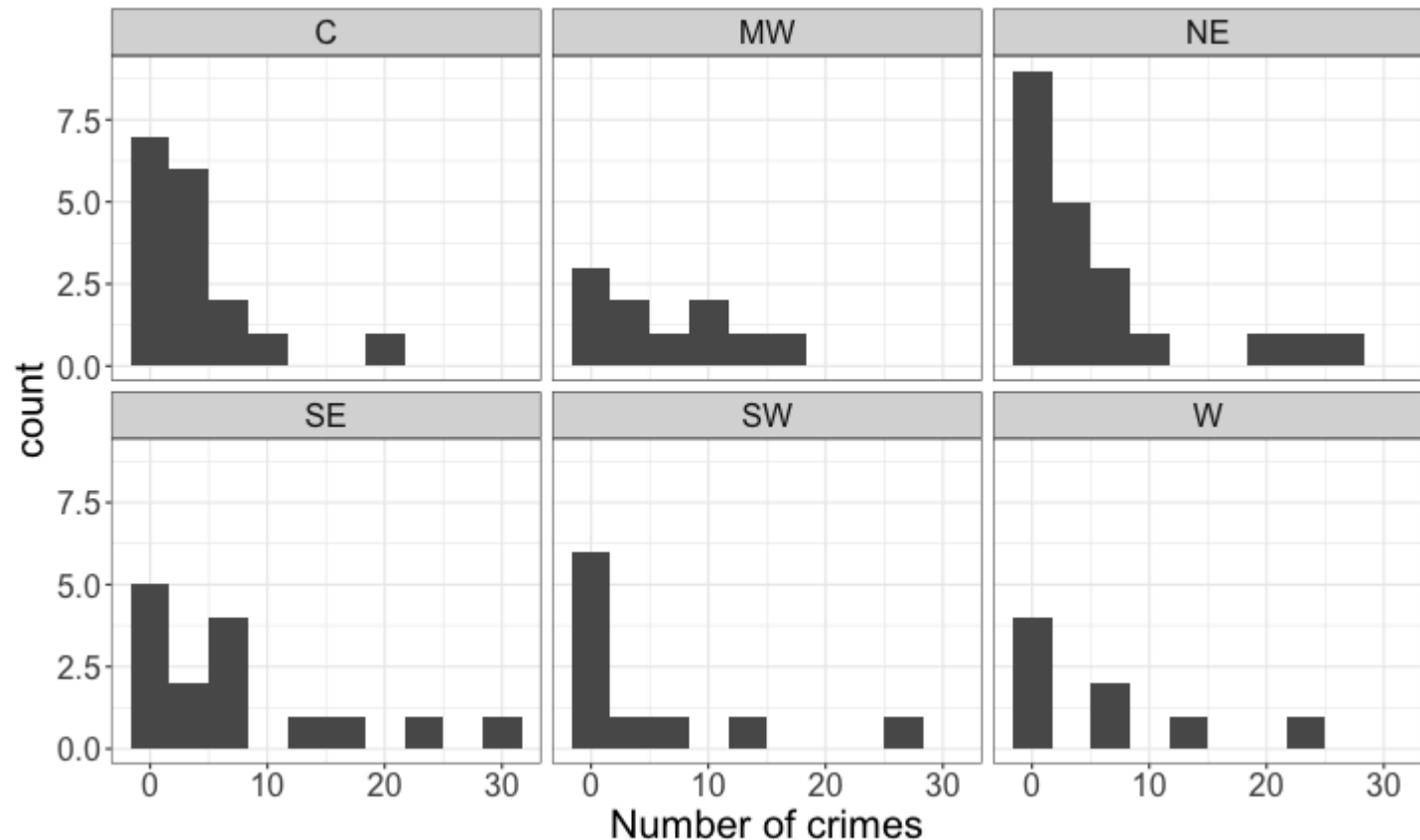> What would be a reasonable model to investigate this question?

Model

$$Crimes_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i$$

What assumptions is this model making?

Exploratory data analysis

Does it look reasonable to assume a Poisson distribution for the response?

Exploratory data analysis

```
crimes %>%
  group_by(region) %>%
  summarize(mean_crimes = mean(nv),
            var_crimes = var(nv))
```

```
## # A tibble: 6 × 3
##    region mean_crimes var_crimes
##    <chr>        <dbl>      <dbl>
## 1 C             3.82       24.3
## 2 MW            6.2        37.1
## 3 NE            5.95       59.0
## 4 SE            8.27       84.4
## 5 SW            5.3        75.3
## 6 W             6.5        65.7
```

Does the Poisson distribution still seem reasonable?

Quantile residual plot

```
m1 <- glm(nv ~ region, data = crimes, family = poisson)
```
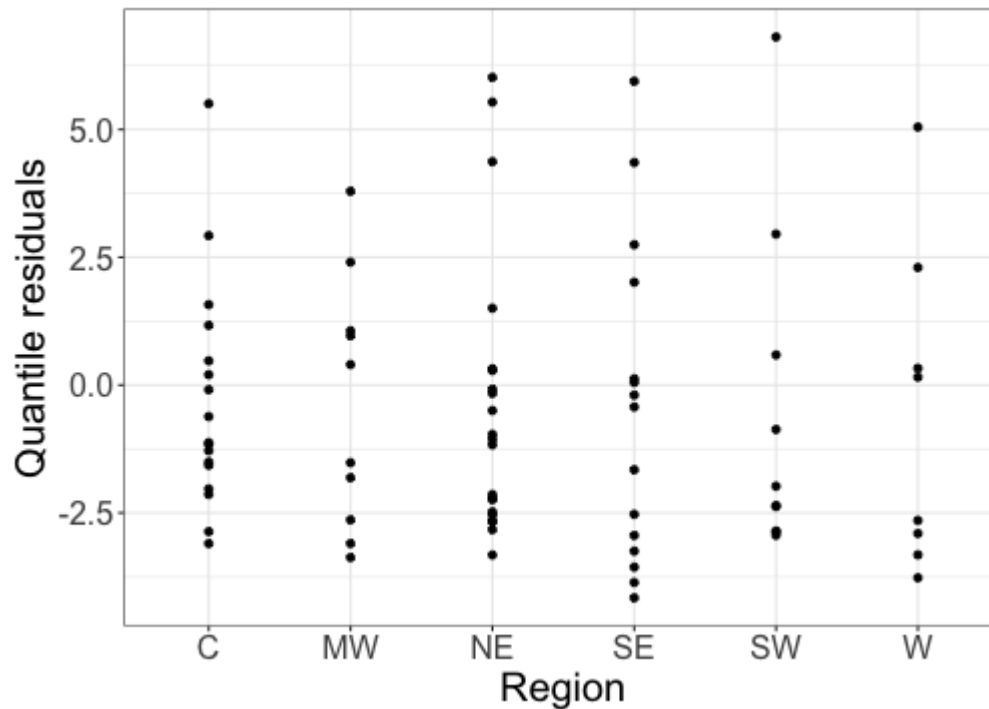
Goodness of fit

Another way to assess whether our model is reasonable is with a *goodness of fit* test.

Goodness of fit test: If the model is a good fit for the data, then the residual deviance follows a χ^2 distribution with the same degrees of freedom as the residual deviance

```
...
##      Null deviance: 649.34   on 80   degrees of freedom
## Residual deviance: 621.24   on 75   degrees of freedom
...
```

Residual deviance = 621.24, df = 75

How likely is a residual deviance of 621.24 if our model is correct?

Goodness of fit

Goodness of fit test: If the model is a good fit for the data, then the residual deviance follows a χ^2 distribution with the same degrees of freedom as the residual deviance

Residual deviance = 621.24, df = 75

```
pchisq(621.24, df=75, lower.tail=F)
```

```
## [1] 5.844298e-87
```

So our model might not be a very good fit to the data.

> Why might our model not be a good fit?

Potential issues with our model

- The Poisson distribution might not be a good choice
- There may be additional factors related to the number of crimes which we are not including in the model

Which other factors might be related to the number of crimes?

Offsets

We will account for school size by including an **offset** in the model:

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i$$
$$+ \log(Enrollment_i)$$

Motivation for offsets

We can rewrite our regression model with the offset:

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i \\ + \log(Enrollment_i)$$

Fitting a model with an offset

```
m2 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = poisson)
summary(m2)
```

```
...
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.30445    0.12403 -10.517  < 2e-16 ***
## regionMW      0.09754    0.17752   0.549  0.58270
## regionNE      0.76268    0.15292   4.987 6.12e-07 ***
## regionSE      0.87237    0.15313   5.697 1.22e-08 ***
## regionSW      0.50708    0.18507   2.740  0.00615 **
## regionW       0.20934    0.18605   1.125  0.26053
...
```

✚ The offset doesn't show up in the output (because we're not estimating a coefficient for it)

Fitting a model with an offset

$$\log(\widehat{\lambda}_i) = -1.30 + 0.10MW_i + 0.76NE_i +$$
$$0.87SE_i + 0.51SW_i + 0.21W_i$$
$$+ \log(Enrollment_i)$$

How would I interpret the intercept -1.30?

When to use offsets

Offsets are useful in Poisson regression when our counts come from groups of very different sizes (e.g., different numbers of students on a college campus). The offset lets us interpret model coefficients in terms of rates instead of raw counts.

> With your neighbor, brainstorm some other data scenarios where our response is a count variable, and an offset would be useful. What would our offset be?