# STATS 3001 / STATS 4104 / STATS 7054
# Statistical Modelling III
# Practical 5 - Poisson regression

Week 9

### GOAL

Perform a Poisson regression to predict the number of people in a househouse based on the age of the head of the household.

### DATA

The Philippine Statistics Authority (PSA) spearheads the Family Income and Expenditure Survey (FIES) nationwide. The survey, which is undertaken every three years, is aimed at providing data on family income and expenditure, including levels of consumption by item of expenditure. The data, from the 2015 FIES, is a subset of 1500 of the 40,000 observations (Philippine Statistics Authority 2015). The data set focuses on five regions: Central Luzon, Metro Manila, Ilocos, Davao, and Visayas.

The data is in the file `fHH1.csv`. Each row is a household, and the follow variables are recorded:

- `location`: where the house is located (Central Luzon, Davao Region, Ilocos Region, Metro Manila, or Visayas)
- `age`: the age of the head of household
- `total`: the number of people in the household other than the head
- `numLT5`: the number in the household under 5 years of age
- `roof`: the type of roof in the household (either Predominantly Light/Salvaged Material, or Predominantly Strong Material.

### STEPS

1. Read in the dataset.

```r
household <- read_csv(here::here("data", "fHH1.csv"))
```

```
## New names:
## * ‘‘ -> ...1
```

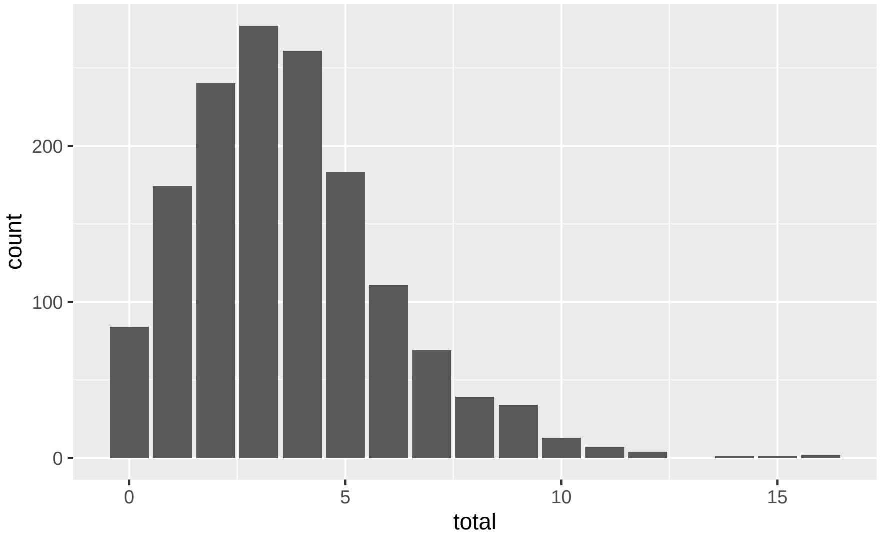
```
## Rows: 1500 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): location, roof
## dbl (4): ...1, age, total, numLT5
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
household
```

```
## # A tibble: 1,500 x 6
##     ...1 location       age total numLT5 roof
##    <dbl> <chr>        <dbl> <dbl>  <dbl> <chr>
##  1     1 CentralLuzon    65     0      0 Predominantly Strong Material
##  2     2 MetroManila     75     3      0 Predominantly Strong Material
##  3     3 DavaoRegion     54     4      0 Predominantly Strong Material
##  4     4 Visayas         49     3      0 Predominantly Strong Material
##  5     5 MetroManila     74     3      0 Predominantly Strong Material
##  6     6 Visayas         59     6      0 Predominantly Strong Material
##  7     7 MetroManila     54     5      0 Predominantly Strong Material
##  8     8 Visayas         41     5      0 Predominantly Strong Material
##  9     9 Visayas         50     6      0 Predominantly Strong Material
## 10    10 CentralLuzon    59     4      0 Predominantly Strong Material
## # ... with 1,490 more rows
```

2. Produce a bar-chart of `total`

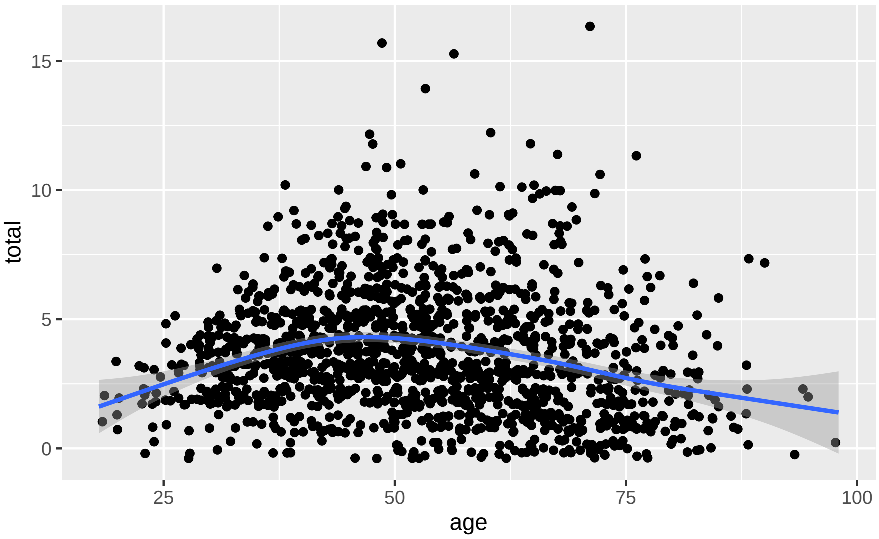
```
household %>%
  ggplot(aes(total)) +
  geom_bar()
```



3. Produce a scatter-plot of `total` against `age` - add a smoothing line.

```
household %>%
  ggplot(aes(age, total)) +
  geom_jitter() +
  geom_smooth()
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



4. Fit the Poisson regression

$$total \sim age$$

```
M1 <- glm(total ~ age, data = household, family = poisson)
summary(M1)
```

```
##
## Call:
## glm(formula = total ~ age, family = poisson, data = household)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9079  -0.9637  -0.2155   0.6092   4.9561
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.5499422  0.0502754  30.829  < 2e-16 ***
## age         -0.0047059  0.0009363  -5.026 5.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2362.5  on 1499  degrees of freedom
## Residual deviance: 2337.1  on 1498  degrees of freedom
## AIC: 6714
##
## Number of Fisher Scoring iterations: 5
```

5. Interpret the coefficient of age.
   The coefficient is $-0.0047059$, so this equates to a decrease of

```
exp(-0.0047059)
```

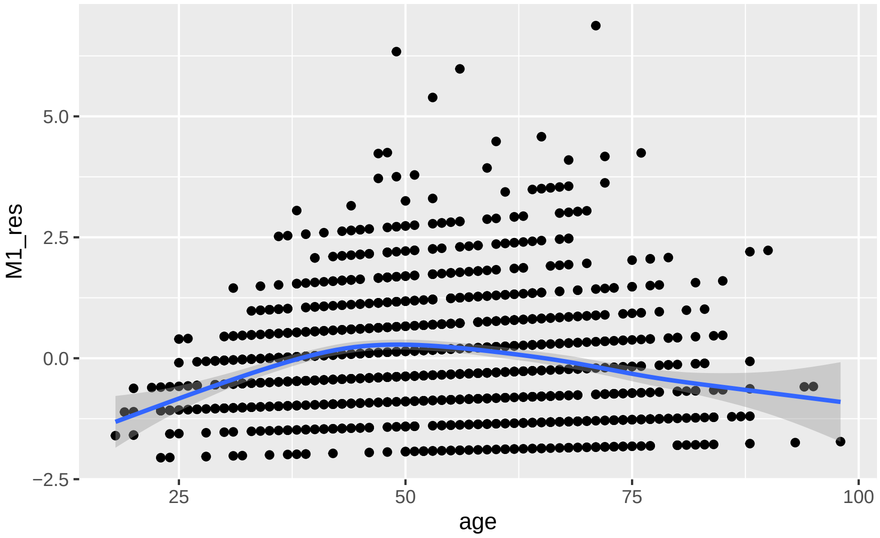
```
## [1] 0.9953052
```

1% for each increase in age of 1 year.
6. Obtain the Pearson residuals. Plot these against `age`. Is the model adequate?

```
household <-
  household %>%
  add_column(
    M1_res = residuals(M1, type = "pearson")
    )
```

```
household %>%
  ggplot(aes(age, M1_res)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



We see a curved pattern to the residuals, and so need to add extra terms to the model.
7. Fit the Poisson regression

$$total \sim age + age^2$$

```
M2 <- glm(total ~ age + I(age^2), data = household, family = poisson)
summary(M2)
```

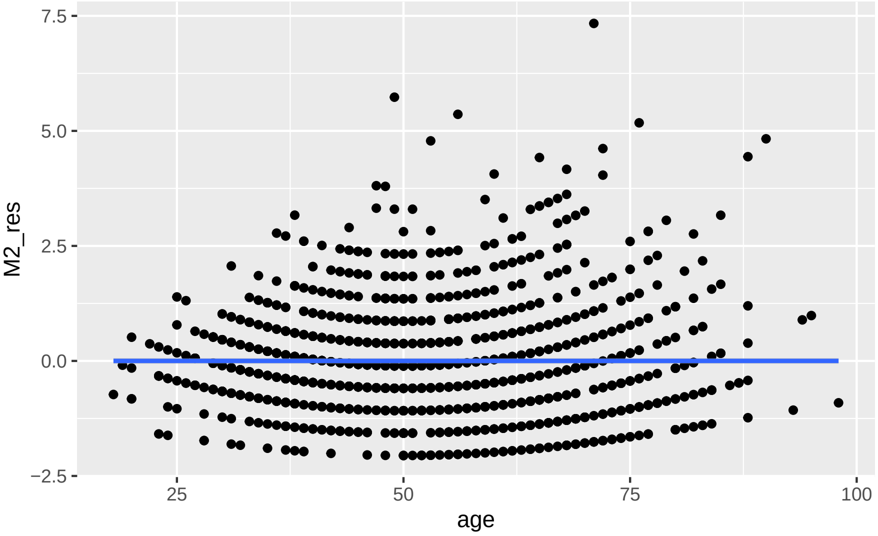
```
##
## Call:
## glm(formula = total ~ age + I(age^2), family = poisson, data = household)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9068  -0.9261  -0.1048   0.5773   5.1731
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.325e-01  1.788e-01  -1.859    0.063 .
## age           7.089e-02  6.890e-03  10.288   <2e-16 ***
## I(age^2)     -7.083e-04  6.406e-05 -11.058   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2362.5  on 1499  degrees of freedom
## Residual deviance: 2200.9  on 1497  degrees of freedom
## AIC: 6579.8
##
## Number of Fisher Scoring iterations: 5
```

8. Repeat the residual plots for the new model.

```
household <-
  household %>%
  add_column(
    M2_res = residuals(M2, type = "pearson")
    )
```

```
household %>%
  ggplot(aes(age, M2_res)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



9. Compare the models using a likelihood ratio test, and AIC.

```
anova(M1, M2, test = "LRT")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: total ~ age
## Model 2: total ~ age + I(age^2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1498     2337.1
## 2      1497     2200.9  1   136.15 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
AIC(M1, M2)
```

```
##    df      AIC
## M1  2 6713.968
## M2  3 6579.823
```

There is a statistically significant difference between the two models and so we use the larger model. Also the AIC is better for the larger model and so should be used. 10. Calculate the predicted values for model M2. What is the age of the head of the household associated with the largest fitted value?

```
household %>%
  add_column(
    fit = predict(M2, type = "response")
  ) %>%
  select(age, fit) %>%
  distinct() %>%
  filter(fit == max(fit))
```

```
## # A tibble: 1 x 2
##     age   fit
##   <dbl> <dbl>
## 1    50  4.22
```

The largest size is 4.22 with the age of the head of the household at 50 years.