glm

Siqi Wang, Yubin Lyu, Liting Wang, Han Xu, JIahao Mao

Introduction

Some introductions

Data Processing

Data Summary

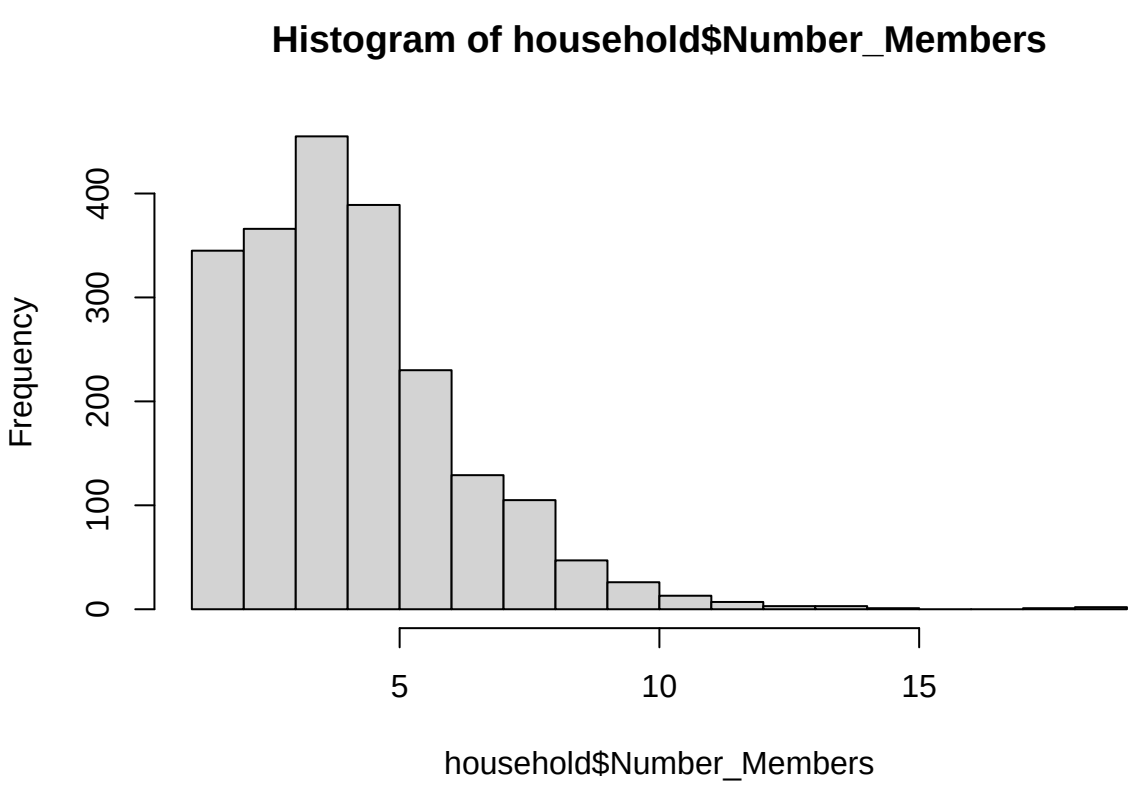
```
 Electricity Householder_Sex                                  Household_Type
 0: 363      Female: 362     Extended Family                         : 585
 1:1759      Male  :1760     Single Family                           :1531
                             Two or More Nonrelated Persons/Members:   6
```

Table 1: Summary statistics of variables

Variable	n	Mean	SD	Min	Median	Max	IQR
Income	2122	1.8e+05	2.3e+05	1.5e+04	1.2e+05	3.2e+06	7.4e+04
FoodExp	2122	7.2e+04	4.5e+04	7.8e+03	6.3e+04	7.3e+05	2.4e+04
Householder_Age	2122	4.9e+01	1.4e+01	9.0e+00	4.8e+01	9.9e+01	1.1e+01
Number_Members	2122	4.5e+00	2.2e+00	1.0e+00	4.0e+00	1.9e+01	2.0e+00
Floorarea	2122	3.6e+01	3.5e+01	5.0e+00	2.6e+01	4.5e+02	1.4e+01
House.Age	2122	1.6e+01	1.1e+01	0.0e+00	1.4e+01	7.5e+01	7.0e+00
Number_bedrooms	2122	1.8e+00	1.0e+00	0.0e+00	2.0e+00	7.0e+00	0.0e+00

Distribution Check

test if the distribution of y is poisson dist



check the skewness and kurtosis results

```
[1] 1.1
```

```
[1] 6.1
```

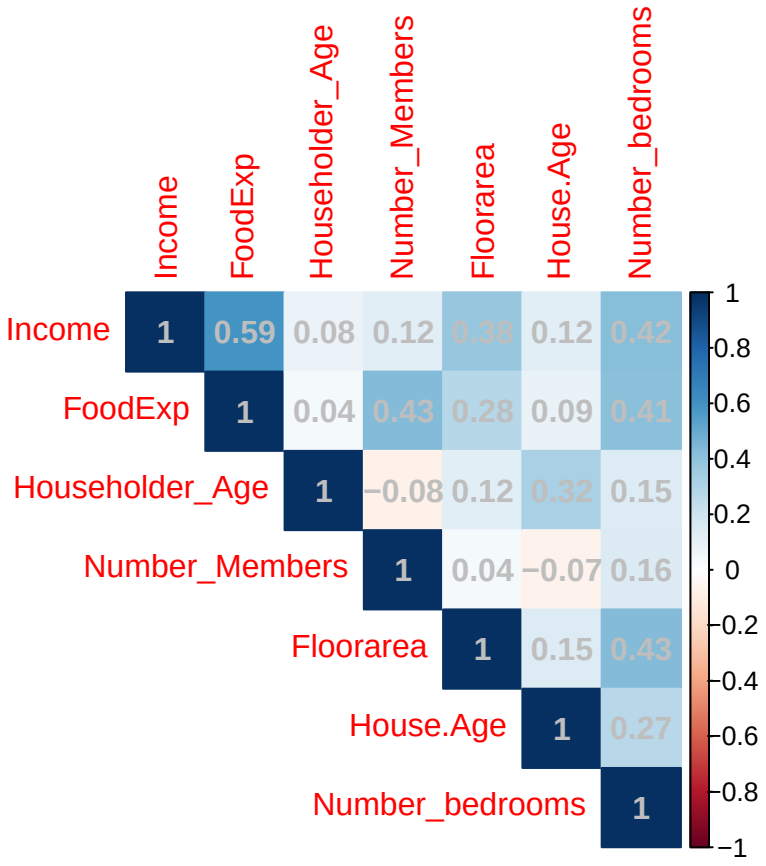
Based on the skewness and kurtosis results, we can determine that the distribution of "y" does not conform to the assumption of a strict Poisson distribution. Specifically, skewness values greater than 1 indicate that the data distribution is right-skewed, and kurtosis values greater than 3 indicate that the data distribution is sharper than the Poisson distribution. In such cases, a Negative Binomial Distribution (NBD) regression model may be considered, as it can be fitted when a Poisson regression model is not up to the task.

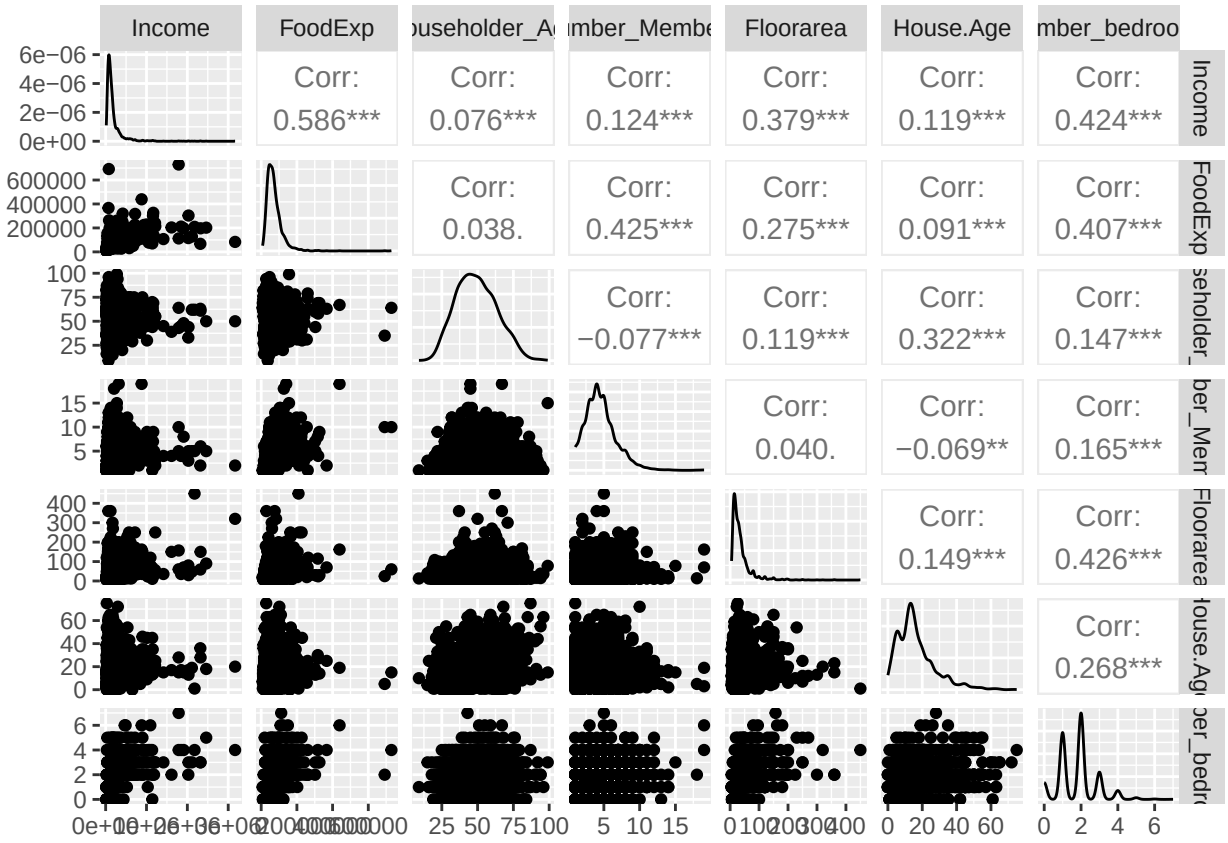
Correlation Matrix and GGpairs

	Income	FoodExp	Householder_Age	Number_Members	Floorarea
Income	1.00	0.59	0.08	0.12	0.38
FoodExp	0.59	1.00	0.04	0.43	0.28
Householder_Age	0.08	0.04	1.00	-0.08	0.12
Number_Members	0.12	0.43	-0.08	1.00	0.04
Floorarea	0.38	0.28	0.12	0.04	1.00
House.Age	0.12	0.09	0.32	-0.07	0.15
Number_bedrooms	0.42	0.41	0.15	0.16	0.43

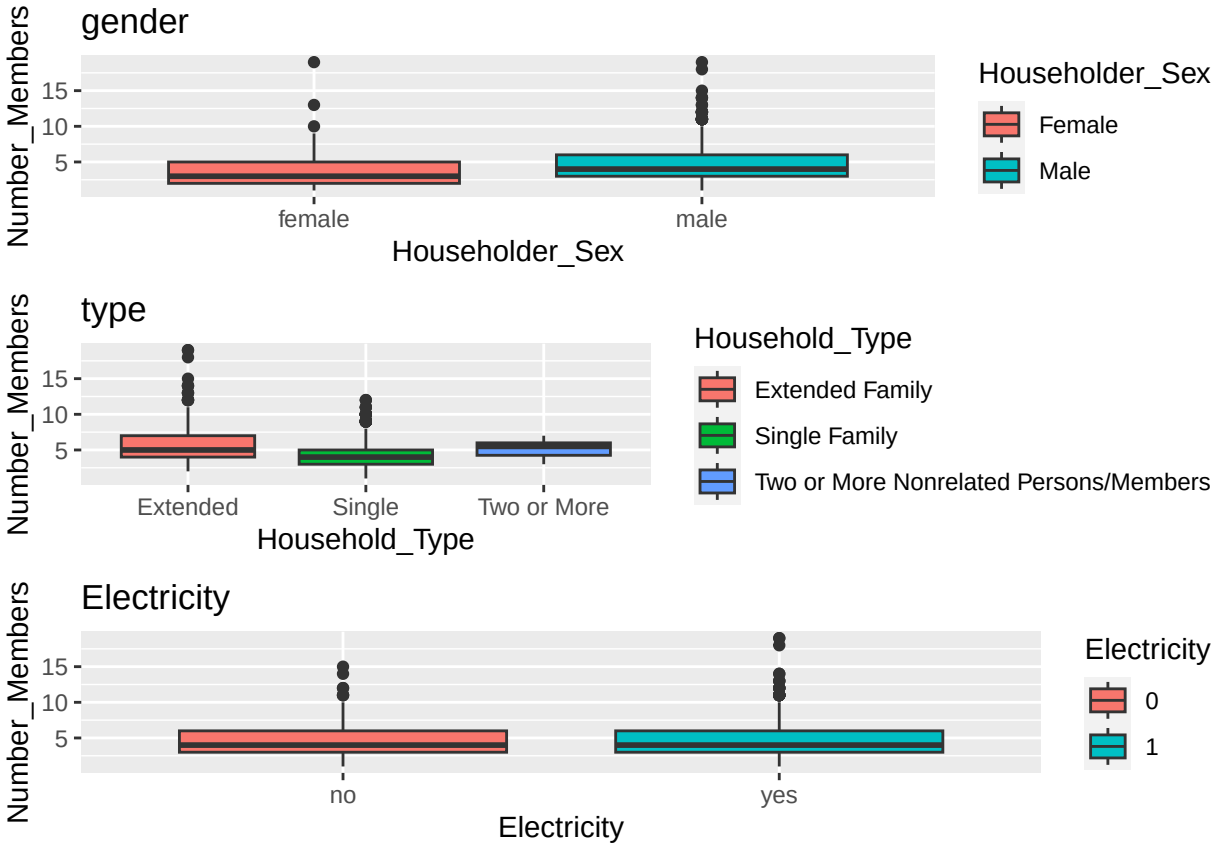
	House.Age	Number_bedrooms
Income	0.12	0.42

```
FoodExp              0.09            0.41
Householder_Age      0.32            0.15
Number_Members      -0.07            0.16
Floorarea            0.15            0.43
House.Age            1.00            0.27
Number_bedrooms      0.27            1.00
```



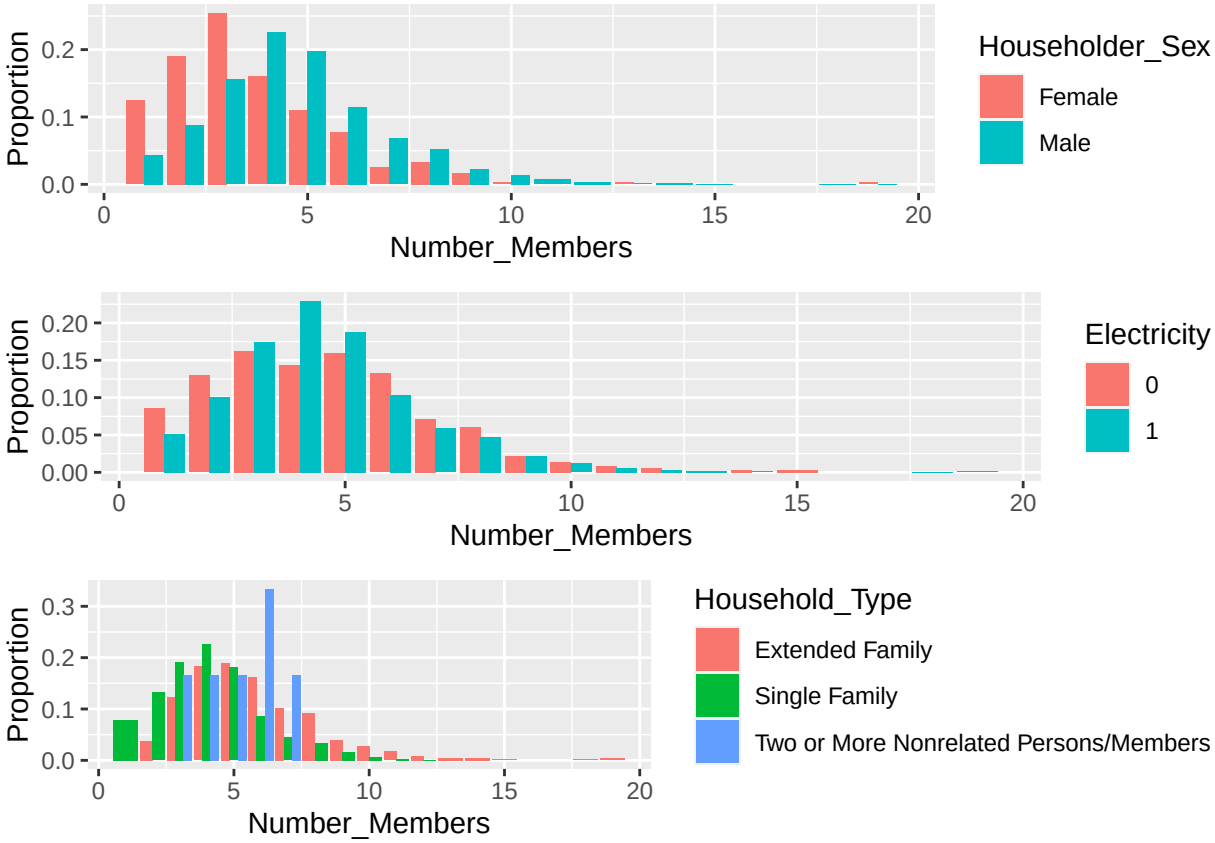


Boxplots of Variables



Histogram of Variables

Number_Members	Female	Male
1	37.2% (45)	62.8% (76)
2	30.8% (69)	69.2% (155)
3	25.1% (92)	74.9% (274)
4	12.7% (58)	87.3% (397)
5	10.3% (40)	89.7% (349)
6	12.2% (28)	87.8% (202)
7	7.0% (9)	93.0% (120)
8	11.4% (12)	88.6% (93)
9	12.8% (6)	87.2% (41)
10	3.8% (1)	96.2% (25)
11	0.0% (0)	100.0% (13)
12	0.0% (0)	100.0% (7)
13	33.3% (1)	66.7% (2)
14	0.0% (0)	100.0% (3)
15	0.0% (0)	100.0% (1)
18	0.0% (0)	100.0% (1)
19	50.0% (1)	50.0% (1)



Model Fitting

```
Call:
glm(formula = Number_Members ~ Income + FoodExp + Householder_Sex +
    Householder_Age + Household_Type + Floorarea + House.Age +
    Number_bedrooms + Electricity, family = poisson(link = "log"),
    data = household)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.523  -0.615  -0.113   0.423   4.115

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                         1.60e+00   6.09e-02
Income                                             -2.39e-07   5.63e-08
FoodExp                                             2.93e-06   1.88e-07
Householder_SexMale                                 2.63e-01   3.05e-02
Householder_Age                                    -3.80e-03   8.10e-04
Household_TypeSingle Family                        -3.47e-01   2.29e-02
Household_TypeTwo or More Nonrelated Persons/Members -1.06e-01   1.81e-01
Floorarea                                          -4.94e-04   3.40e-04
House.Age                                          -3.71e-03   1.03e-03
```

```
Number_bedrooms                                          5.01e-02   1.23e-02
Electricity1                                            -9.03e-02   2.85e-02
                                                        z value Pr(>|z|)
(Intercept)                                               26.21  < 2e-16 ***
Income                                                    -4.23  2.3e-05 ***
FoodExp                                                   15.59  < 2e-16 ***
Householder_SexMale                                        8.62  < 2e-16 ***
Householder_Age                                           -4.68  2.8e-06 ***
Household_TypeSingle Family                              -15.13  < 2e-16 ***
Household_TypeTwo or More Nonrelated Persons/Members      -0.59  0.55842
Floorarea                                                 -1.45  0.14648
House.Age                                                 -3.61  0.00031 ***
Number_bedrooms                                            4.06  4.9e-05 ***
Electricity1                                              -3.17  0.00154 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1551.8  on 2111  degrees of freedom
AIC: 8512

Number of Fisher Scoring iterations: 5



Call:
glm(formula = Number_Members ~ log(Income) + log(FoodExp) + Householder_Age +
    Floorarea + House.Age + Number_bedrooms, family = poisson(link = "log"),
    data = household)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.208  -0.614  -0.134   0.449   3.780

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.620147   0.233877  -15.48  < 2e-16 ***
log(Income)     -0.166883   0.020604   -8.10  5.5e-16 ***
log(FoodExp)     0.650278   0.028050   23.18  < 2e-16 ***
Householder_Age -0.001247   0.000792   -1.57   0.1154
Floorarea       -0.001011   0.000343   -2.95   0.0032 **
House.Age       -0.004030   0.001030   -3.91  9.2e-05 ***
Number_bedrooms  0.024973   0.012608    1.98   0.0476 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1508.7  on 2115  degrees of freedom
AIC: 8461

Number of Fisher Scoring iterations: 4
```

Use BIC to do variable selection

```
Call:
bic.glm.formula(f = Number_Members ~ Income + FoodExp + Householder_Sex +     Householder_Age + Household


  5  models were selected
 Best  5  models (cumulative posterior probability =  1 ):

                                                  p!=0    EV
Intercept                                         100     1.58e+00
Income                                            100.0  -2.56e-07
FoodExp                                           100.0   2.95e-06
Householder_Sex                                   100.0
               .Male                                       2.63e-01
Householder_Age                                   100.0  -3.85e-03
Household_Type                                    100.0
              .Single Family                              -3.46e-01
              .Two or More Nonrelated Persons/Members     -1.02e-01
Floorarea                                          4.4   -2.15e-05
House.Age                                         96.4   -3.68e-03
Number_bedrooms                                   96.5    4.17e-02
Electricity                                       76.5
           .1                                             -7.02e-02

nVar
BIC
post prob
                                                  SD        model 1
Intercept                                         6.53e-02   1.60e+00
Income                                            5.65e-08  -2.53e-07
FoodExp                                           1.90e-07   2.93e-06
Householder_Sex
               .Male                              3.05e-02   2.63e-01
Householder_Age                                   8.23e-04  -3.85e-03
Household_Type
              .Single Family                      2.30e-02  -3.47e-01
              .Two or More Nonrelated Persons/Members  1.81e-01  -1.02e-01
Floorarea                                         1.23e-04      .
House.Age                                         1.25e-03  -3.76e-03
Number_bedrooms                                   1.43e-02   4.45e-02
Electricity
           .1                                     4.63e-02  -9.13e-02

nVar                                                         8
BIC                                                      -1.46e+04
post prob                                                 0.685
                                                  model 2    model 3
Intercept                                          1.53e+00   1.60e+00
Income                                            -2.76e-07  -2.39e-07
FoodExp                                            2.96e-06   2.93e-06
Householder_Sex
               .Male                               2.63e-01   2.63e-01
```

```
Householder_Age                                        -3.78e-03  -3.80e-03
Household_Type
             .Single Family                            -3.43e-01  -3.47e-01
             .Two or More Nonrelated Persons/Members   -1.04e-01  -1.06e-01
Floorarea                                                  .      -4.94e-04
House.Age                                              -4.11e-03  -3.71e-03
Number_bedrooms                                         3.88e-02   5.01e-02
Electricity
          .1                                               .      -9.03e-02

nVar                                                       7          9
BIC                                                    -1.46e+04  -1.46e+04
post prob                                               0.201      0.044
                                                       model 4    model 5
Intercept                                               1.59e+00   1.57e+00
Income                                                 -2.55e-07  -2.23e-07
FoodExp                                                 2.97e-06   3.06e-06
Householder_Sex
              .Male                                     2.66e-01   2.61e-01
Householder_Age                                        -4.61e-03  -3.64e-03
Household_Type
             .Single Family                            -3.44e-01  -3.50e-01
             .Two or More Nonrelated Persons/Members   -8.18e-02  -1.11e-01
Floorarea                                                  .          .
House.Age                                                  .      -3.43e-03
Number_bedrooms                                         3.66e-02      .
Electricity
          .1                                           -1.03e-01      .

nVar                                                       7          6
BIC                                                    -1.46e+04  -1.46e+04
post prob                                               0.036      0.035

  1  observations deleted due to missingness.


Call:
glm(formula = Number_Members ~ Income + FoodExp + Householder_Sex +
    Householder_Age + Household_Type + House.Age + Number_bedrooms +
    Electricity, family = "poisson", data = household)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.522  -0.619  -0.110   0.426   4.105

Coefficients:
                                                       Estimate Std. Error
(Intercept)                                            1.60e+00   6.09e-02
Income                                                -2.53e-07   5.54e-08
FoodExp                                                2.93e-06   1.88e-07
Householder_SexMale                                    2.63e-01   3.05e-02
Householder_Age                                       -3.85e-03   8.10e-04
Household_TypeSingle Family                           -3.47e-01   2.29e-02
Household_TypeTwo or More Nonrelated Persons/Members -1.02e-01   1.81e-01
```

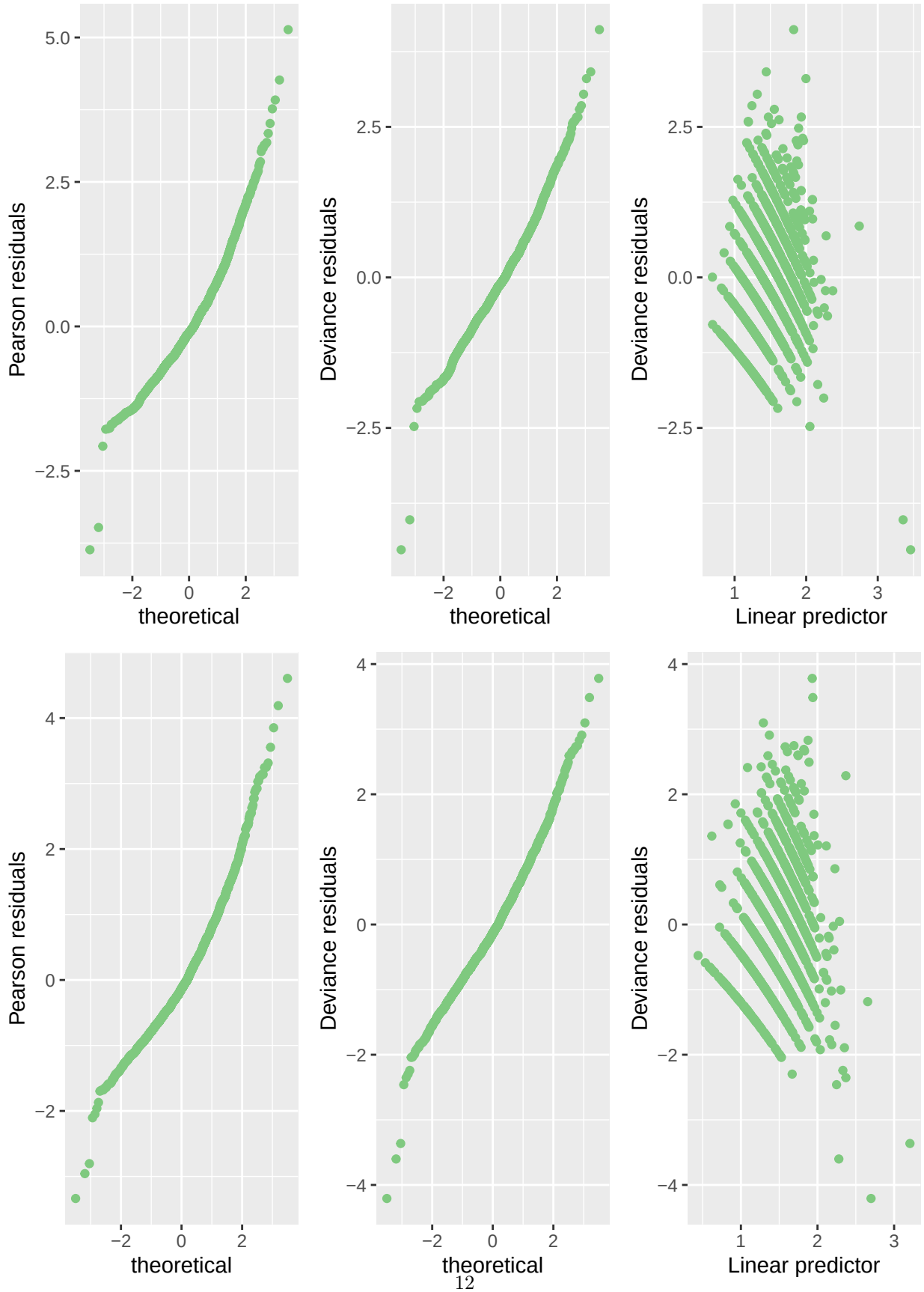
```
House.Age                                            -3.76e-03   1.03e-03
Number_bedrooms                                       4.45e-02   1.17e-02
Electricity1                                         -9.13e-02   2.85e-02
                                                     z value Pr(>|z|)
(Intercept)                                            26.20  < 2e-16 ***
Income                                                 -4.57  4.8e-06 ***
FoodExp                                                15.60  < 2e-16 ***
Householder_SexMale                                     8.63  < 2e-16 ***
Householder_Age                                        -4.76  2.0e-06 ***
Household_TypeSingle Family                           -15.15  < 2e-16 ***
Household_TypeTwo or More Nonrelated Persons/Members   -0.56  0.57331
House.Age                                              -3.65  0.00026 ***
Number_bedrooms                                         3.79  0.00015 ***
Electricity1                                           -3.21  0.00135 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1554.0  on 2112  degrees of freedom
AIC: 8512

Number of Fisher Scoring iterations: 5
```

Negative Binomial Distribution

```
Call:
glm.nb(formula = Number_Members ~ Income + FoodExp + Householder_Sex +
    Householder_Age + Household_Type + Floorarea + House.Age +
    Number_bedrooms + Electricity, data = household, init.theta = 76069.31481,
    link = log)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.523  -0.615  -0.113   0.423   4.114

Coefficients:
                                                      Estimate Std. Error
(Intercept)                                           1.60e+00   6.09e-02
Income                                               -2.39e-07   5.63e-08
FoodExp                                               2.93e-06   1.88e-07
Householder_SexMale                                   2.63e-01   3.05e-02
Householder_Age                                      -3.80e-03   8.11e-04
Household_TypeSingle Family                          -3.47e-01   2.29e-02
Household_TypeTwo or More Nonrelated Persons/Members -1.06e-01   1.81e-01
Floorarea                                            -4.94e-04   3.40e-04
House.Age                                            -3.71e-03   1.03e-03
Number_bedrooms                                       5.01e-02   1.23e-02
Electricity1                                         -9.03e-02   2.85e-02
                                                     z value Pr(>|z|)
(Intercept)                                            26.21  < 2e-16 ***
```

```
Income                                                -4.23  2.3e-05 ***
FoodExp                                               15.59  < 2e-16 ***
Householder_SexMale                                    8.62  < 2e-16 ***
Householder_Age                                       -4.68  2.8e-06 ***
Household_TypeSingle Family                          -15.13  < 2e-16 ***
Household_TypeTwo or More Nonrelated Persons/Members  -0.59  0.55846
Floorarea                                             -1.45  0.14646
House.Age                                             -3.61  0.00031 ***
Number_bedrooms                                        4.06  4.9e-05 ***
Electricity1                                          -3.17  0.00154 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(76069) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1551.7  on 2111  degrees of freedom
AIC: 8514

Number of Fisher Scoring iterations: 1

              Theta:  76069
          Std. Err.:  280723
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -8490
```

Deviance plots



Model Evaluation

```
[1] 1552 8512
```

```
[1] 1509 8461
```

```
[1] 1554 8512
```

```
[1] 1552 8514
```

Goodness-of-fit test

```
Chi-square test statistic =  1584
```

```
df =  2111
```

```
p-value =  1
```

	OR	2.5 %	97.5 %
(Intercept)	4.94	4.38	5.57
Income	1.00	1.00	1.00
FoodExp	1.00	1.00	1.00
Householder_SexMale	1.30	1.23	1.38
Householder_Age	1.00	0.99	1.00
Household_TypeSingle Family	0.71	0.68	0.74
Household_TypeTwo or More Nonrelated Persons/Members	0.90	0.62	1.26
Floorarea	1.00	1.00	1.00
House.Age	1.00	0.99	1.00
Number_bedrooms	1.05	1.03	1.08
Electricity1	0.91	0.86	0.97