Generalized linear models

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

IDA
Intelligent Data Analysis
RESEARCH GROUP

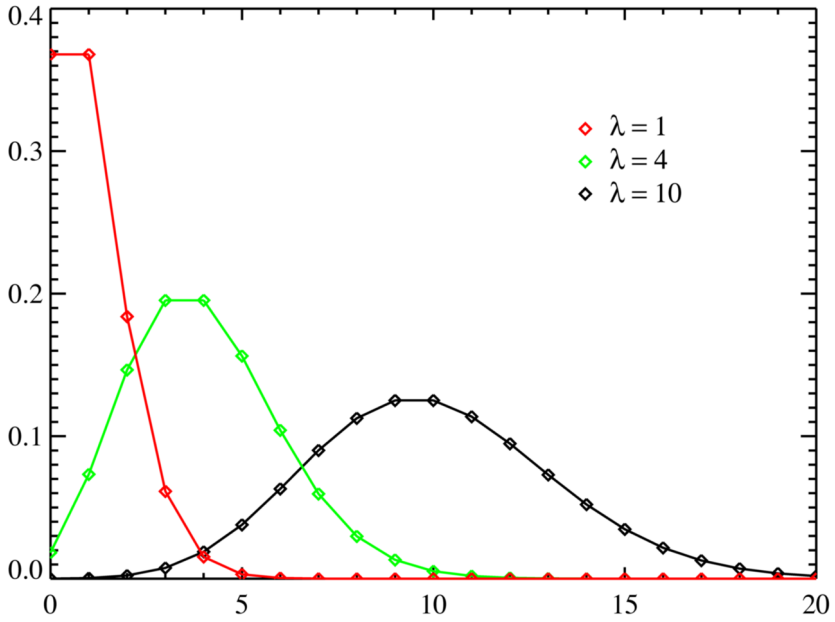
http://cw.felk.cvut.cz/wiki/courses/b4m36san/start

Introduction

- Logistic regression
  - is a linear model too,
  - logit link function was introduced to map between binary outcome and linear predictor,
- A similar approach could be applied to other types of outcome variables
  - Poisson regression as another example will be given,
- **generalized linear models** (GLMs)
  - will eventually cover a whole class of these models,
  - the same principle for the entire family of exponential distributions,
  - GLMs differ in link function and probability distribution
    - the former "links" the linear predictor and the parameters for probability distribution,
    - the latter generates the dependent variable.

Poisson distribution

- Poisson regression assumes the response variable Y has a Poisson distribution for each level of X
  - an event happening a certain number of times (k) within a given interval of time or space,
  - example: machine malfunctions per year, male grizzly bears per hectare,
  - often referred to as count data too.



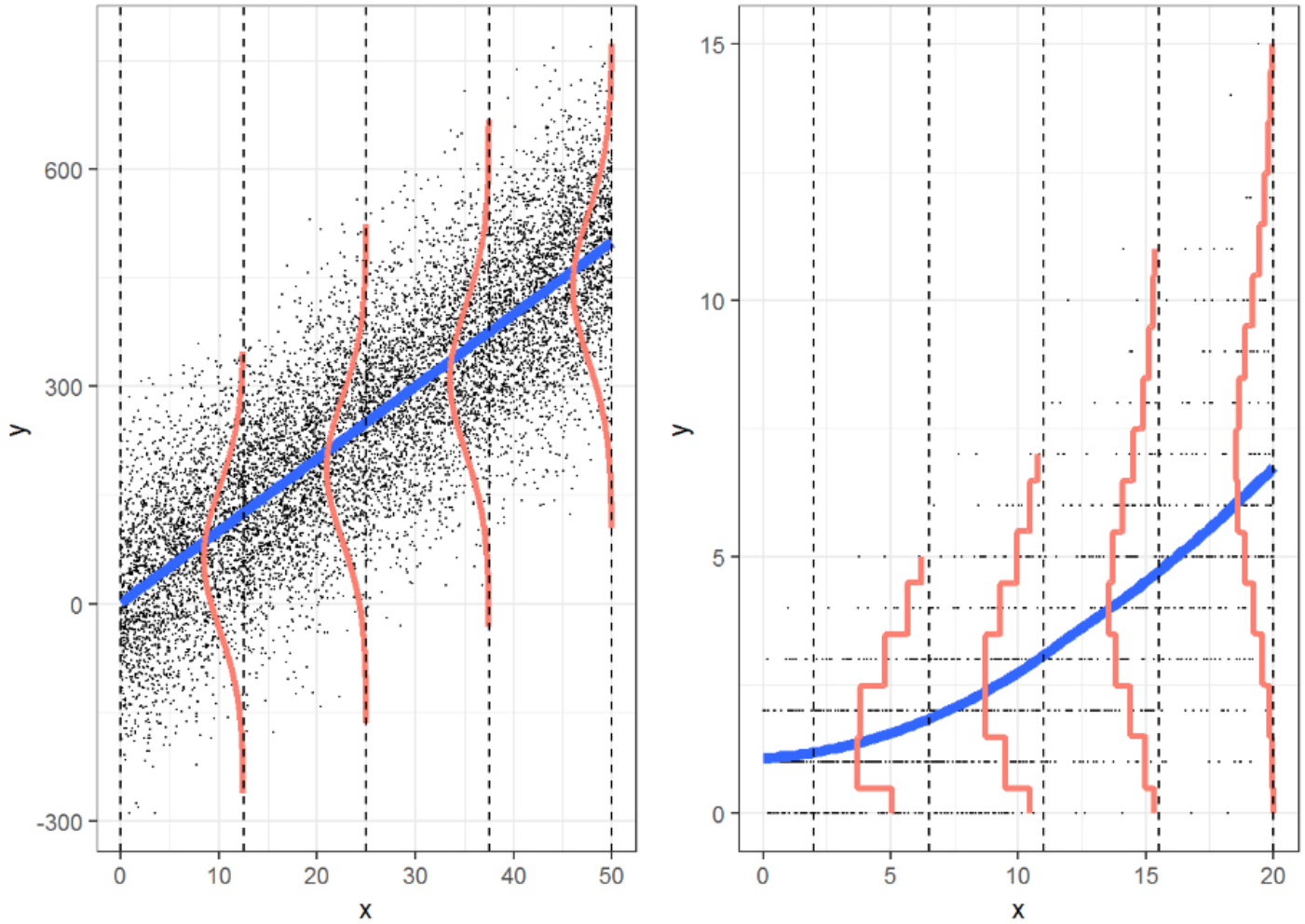
$$\lambda > 0, \lambda = E(Y) = Var(Y)$$

$$Pr(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Can we use linear regression for a Poisson outcome?

- yes, we can
  - for large λs, the Poisson distribution can be approximated with the Normal distribution (Poisson(λ) $\approx$ N($\mu = \lambda$,$\sigma = \sqrt{\lambda}$)),
- however, linear regression
  - can easily predict negative counts,
  - assumes that variance does not change with mean, count data are characterized by heteroscedasticity,
  - assumes that error distribution is not skewed, count data are skewed,
  - assumes linearity between the mean count and the predictors, the relationship can be arbitrary.
- **Poisson regression** is more appropriate.

Linear vs Poisson regression



Roback and Legler: Poisson Regression.

Poisson regression

- Recall that with linear regression

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$
$$y_i \sim \mathcal{N}(\mu_i, \epsilon)$$

- in Poisson regression

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$
$$y_i \sim Poisson(\mu_i)$$

- logistic regression has a similar form

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$
$$q_i = \frac{1}{1 + e^{-\eta_i}}$$
$$y_i \sim Bernoulli(q_i)$$

Anscombe's quartet ...

- What would you say about the following model?

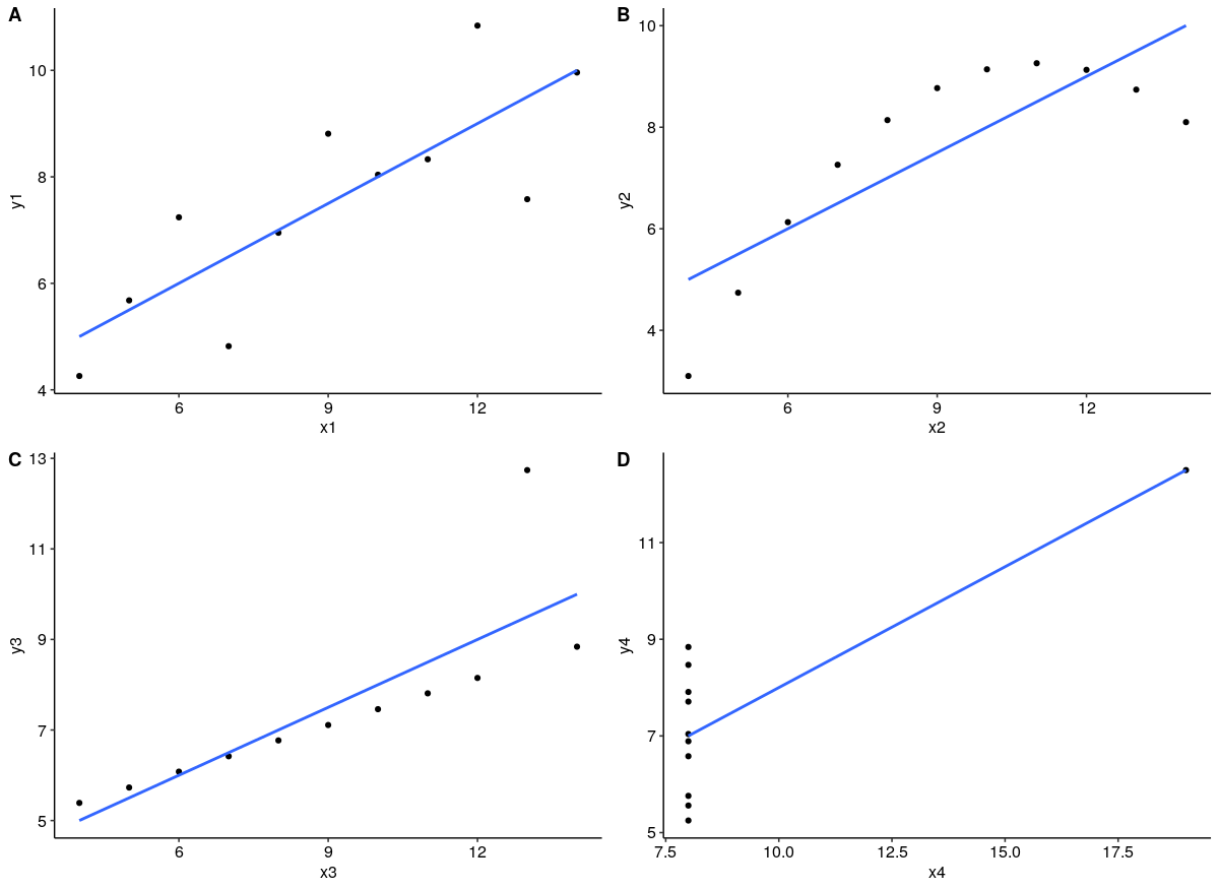
```
summary(lm(y ~ x,d))

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671   0.02559 *
x             0.4999     0.1178   4.243   0.00216 **

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,  Adjusted R-squared:  0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Anscombe's quartet . . .

- What would you say about the following model?



Generalized linear models

- GLM is a flexible generalization of ordinary linear regression,
- it consists of three elements
  - a **linear predictor** $\eta = X^T\beta$,
  - a **link function** g such that $E(Y \mid X) = \mu = g^{-1}(\eta)$,
  - a particular **distribution** for modeling Y from among those which are considered exponential families of probability distributions,
- for previously known regression types
  - linear: identity function + normal distribution,
  - Poisson: log function + Poisson distribution,
  - logistic: logit function + binomial distribution.
- and many other (custom) pairs.

Generalized linear models – exponential family

- GLMs assume that the outcome y has an exponential conditional distribution
  - let us deal with one-parameter distributions only to simplify

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

  - $y \in \mathbb{R}$, ϕ is an known dispersion, θ is the only canonical parameter,
  - the main restriction: y and θ interact only in one multiplicative term,
- easy to show that $Poisson(\mu) = \frac{\mu^y e^{-\mu}}{y!}$ falls in $f_\theta(y)$
  - we only need to know that $\mu^y = \exp(y \log \mu)$,
  - $\frac{\mu^y e^{-\mu}}{y!} = \exp\left(y \log \mu - \mu - \log(y!)\right)$,
  - thus $\theta = \log \mu$, $b(\theta) = \mu$, $c(y, \phi) = -\log(y!)$, $\phi = 1$,
  - also $\mu = e^\theta$, $b(\theta) = e^\theta$, $b'(\theta) = e^\theta$.

Generalized linear models – learning

- GLMs maximize log likelihood to optimize models
  - for exponential family it has a convenient form

$$\ell(\theta) = \log f_\theta(y) = \frac{Y\theta - b(\theta)}{\phi} + c(Y;\phi)$$

  - as it holds $\int f_\theta(y)\,dy \equiv 1$ we can also use

$$\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) = 0$$

  - therefore we may assume

$$\frac{\partial \ell}{\partial \theta} = \frac{Y - b'(\theta)}{\phi} \rightarrow \mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) = \frac{\mathbb{E}(Y) - b'(\theta)}{\phi}$$

$$\mathbb{E}(Y) = \mu = b'(\theta)$$

Generalized linear models – learning

- However, we have to optimize parameters β not θ
  - β relationship to θ is mediated through link function g

$$g(\mu) = X^T\beta$$

  - g can be an arbitrary monotone increasing and differentiable function

$$\mu = g^{-1}(X^T\beta)$$

  - still, it is convenient, if we choose the **canonical** link function, so that

$$g(\mu) = \theta$$

  - given $\mu = b'(\theta)$ it implies that

$$g(\mu) = (b')^{-1}(\mu)$$

Generalized linear models – learning

- Maximum likelihood estimation with a general link function g

$$\ell_n(\beta; \mathbf{Y}, \mathbf{X}) = \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} = \sum_i \frac{Y_i(g \circ b')^{-1}(X_i^T\beta) - b((g \circ b')^{-1}(X_i^T\beta))}{\phi}$$

- maximum likelihood estimation with the canonical link function g

$$\ell_n(\beta; \mathbf{Y}, \mathbf{X}) = \sum_i \frac{Y_i X_i^T\beta - b(X_i^T\beta)}{\phi}$$

  - $\ell(\theta)$ is strictly concave (given $\phi > 0$),
  - as a consequence the ML estimator is unique,

- in Poisson regression we have already shown that $b(\theta) = e^\theta$ and thus the canonical link for this family must be

$$g(\mu) = (b')^{-1}(\mu) \to g(\mu) = \log(\mu)$$

Linear models – evaluation and comparisons

- We have already seen than linear models can be compared with F tests
  - we compared our model with the intercept-only model to test whether at least one predictor in our model is useful

$$F = \frac{(TSS - RSS)/p}{RSS/(m - p - 1)}$$

  - m is the sample size, p is the number of predictors, TSS quantifies the error of the intercept-only model, RSS quantifies the error of our model,
- the formula could be generalized to compare a pair of **nested** linear models
  - model$_1$ has p_1 predictors that **make a subset** of p_2 predictors in model$_2$

$$F = \frac{(RSS_1 - RSS_2)/(df_1 - df_2)}{RSS_2/df_2}$$

Linear models – evaluation and comparisons

- Since $df_1 = m - p_1 - 1$ and $df_2 = m - p_2 - 1$ it also holds

$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(m - p_2 - 1)}$$

- in R the test can be performed with `anova()`
  - `anova(lm(...),lm(...))` for a pair of models,
  - or `anova(lm(...),lm(...),...,lm(...))` for multiple models,
  - where size of the models grows and adjoining models are compared,
- the nested models could also be tested with `aov()`
  - where `summary(aov(lm(y ~ x1 + ... + xp,d)))` does the same as
  - `anova(lm(y ~ 1,d)),lm(y ~ x1,d)),...,lm(y ~ x1 + ... + xp,d))`.

Generalized linear models – evaluation and comparisons

- The same principle for GLMs, variance replaced by **deviance**
  - it relates log likelihoods of our model (θ_m) and the saturated model (θ_s)

$$D(\theta_m) = 2(\ell(\theta_s; \mathbf{Y}, \mathbf{X}) - \ell(\theta_m; \mathbf{Y}, \mathbf{X}))$$

  - saturated model fits the data perfectly
    - it has as many parameters as samples,
    - it does not have to have zero log likelihood anyway,
  - the smaller the deviance, the better the model,
- eventually, deviances of two (or more) nested models can be compared with
  - `anova(glm(...),...,glm(...),test="LRT")`,
  - which is (a series of) the likelihood ratio test(s)
    - H_0: both the (adjoining) models fit the data equally well,
    - H_a: the larger model significantly outperforms the nested model,

Generalized linear models – evaluation and comparisons

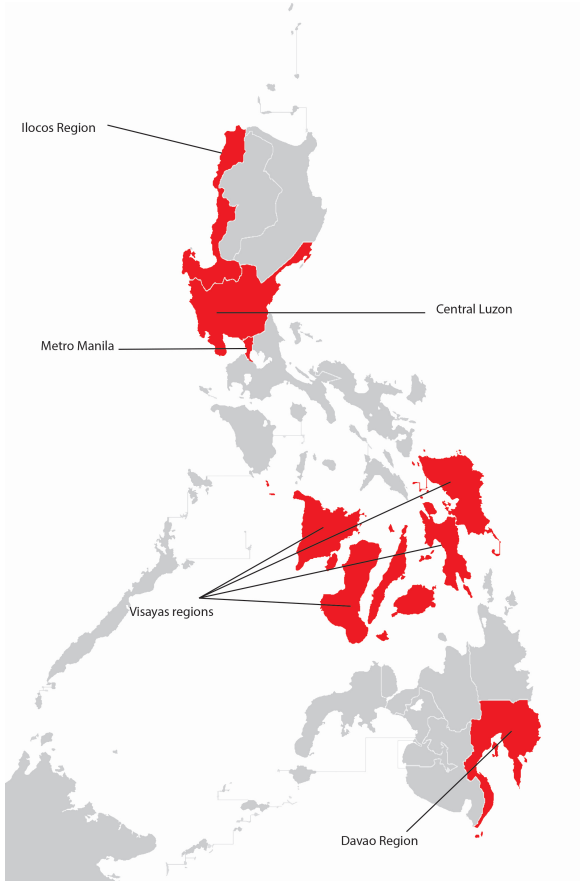
- If a pair of models is not nested deviances could be misleading
  - saturated models may change (e.g., if we change GLM family),
  - different parametric spaces make likelihood ratio tests impossible,
- these models can be compared e.g. in terms of their AIC
  - Akaike information criterion

$$AIC(\theta; \mathbf{Y}, \mathbf{X}) = 2p - 2\ell(\theta; \mathbf{Y}, \mathbf{X})$$

  - AIC is a minimization criterion,
  - an estimator of out-of-sample prediction error,
  - a means for model selection ($\mathbf{Y}$ and $\mathbf{X}$ must be kept unchanged),
- to relate non-nested models in R
  - use `compareGLM(glm(...),...,glm(...))`,
  - calculates more quality measures (AIC, BIC, etc.).

Example: Household Size in the Philippines

- The goal: predict the number of people sharing a house as a function of the age of the household head and location/island,
- the dataset: 1,500 households, three variables of interest (age, location, total).

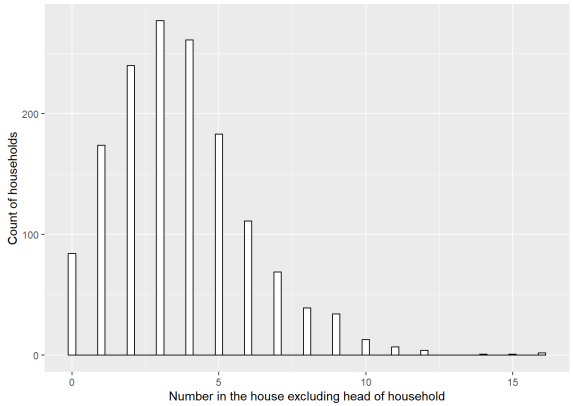


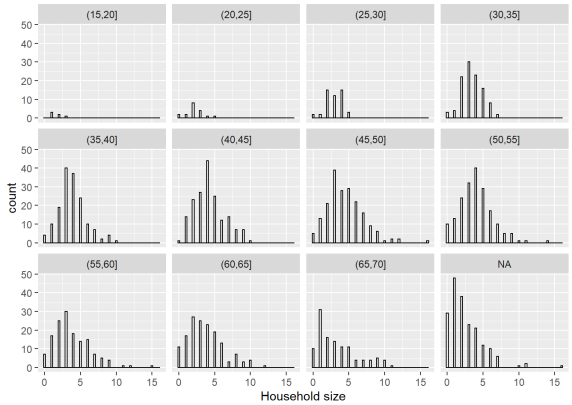
The first five observations from the Philippines Household case study.

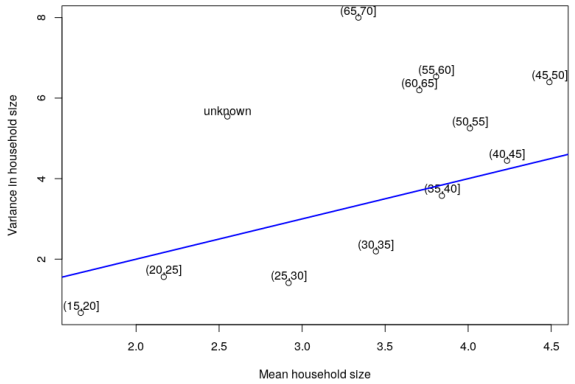
X1	location	age	total	numLT5	roof
1	CentralLuzon	65	0	0	Predominantly Strong Material
2	MetroManila	75	3	0	Predominantly Strong Material
3	DavaoRegion	54	4	0	Predominantly Strong Material
4	Visayas	49	3	0	Predominantly Strong Material
5	MetroManila	74	3	0	Predominantly Strong Material

Example: Household Size in the Philippines

- Let us start with exploratory data analysis (EDA)
  - in order to propose the right GLM,
- let us plot the distribution of the target variable and also its distribution conditioned by age
  - conclusion #1: count target variable whose mean is influenced by age and can be modelled with a Poisson distribution ($\mu = E(total) \approx Var(total)$).

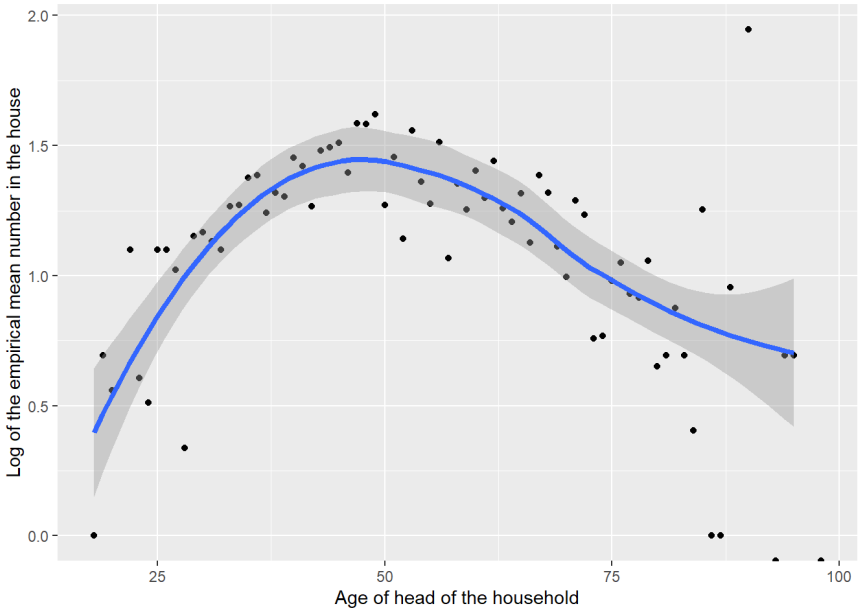


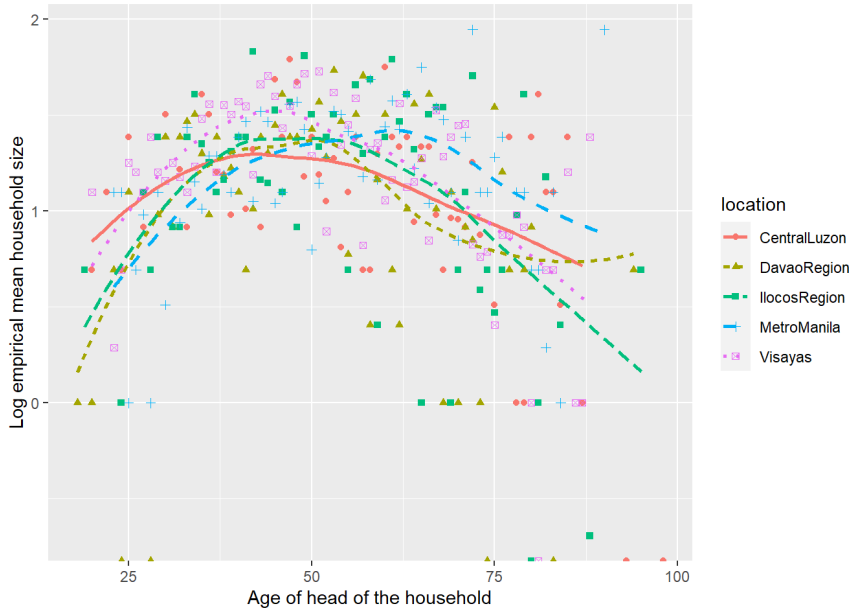




Example: Household Size in the Philippines

- The canonical link function for Poisson regression is log function
  - is the relationship between age and the household size exponential?
  - conclusion #2: a different link or age non-linear transformations needed,
- Location does not influence the shape of age vs household size relationship
  - conclusion #3: no location:age interaction term needed.





Example: Household Size in the Philippines

- Based on EDA: $\log(\mu) = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{age}^2 + \beta_3 \times \text{location}$
- let us construct the model in R

```
m.full <- glm(total ~ age + age2 + location,
              family = poisson, data = fHH1)
coef(summary(m.full))
```

	Estimate	Std. error	z value	p-value
(Intercept)	-0.3843	1.821e-01	-2.111	3.480e-02
age	0.0704	6.905e-03	10.19	2.197e-24
age2	-0.0007	6.420e-05	-10.94	7.126e-28
locDavao	-0.0194	5.378e-02	-0.360	7.185e-01
locIlocos	0.0610	5.266e-02	1.158	2.468e-01
locManila	0.0545	4.720e-02	1.154	2.484e-01
locVisayas	0.1121	4.175e-02	2.685	7.247e-03

Example: Household Size in the Philippines

- Let us check how our model works (in comparison with alternative models)
  - we will test the drop in deviance in nested models.

```
m.null <- glm(total ~ 1, family = poisson, data = fHH1)
m.age <- glm(total ~ age, family = poisson, data = fHH1)
m.age2 <- glm(total ~ age+age2, family = poisson, data = fHH1)
anova(m.null,m.age,m.age2,m.full,test = "Chisq")
```

```
          ResidDf   Resid   DevDf Deviance   Pr(>Chi)
m.null       1499  2362.5
m.age        1498  2337.1    1     25.399  4.661e-07 ***
m.age2       1497  2200.9    1    136.145  < 2.2e-16 ***
m.full       1493  2187.8    4     13.144    0.01059 *
```

- Is Poisson regression helpful?
  - it clearly is as `AIC(m.full)=6575 < AIC(lm.full)=6731`.

Summary

- GLM is a broader class of models that generalizes multiple linear regression
  - all GLMs have similar forms for their likelihoods, MLEs, and deviances,
  - easier to find model estimates and their corresponding uncertainty,
  - OLS (ordinary least squares) replaced by IRLS (iteratively reweighted least squares),
- assumptions less strict than in multiple linear regression
  - observations still must be independent,
  - the distribution of residuals can be from the exponential family,
  - the homogeneity of variance does not need to be satisfied,
- GAM is a more recent concept emphasizing non-linear transformations
  - as we could see, non-linear transformations can be aplied in GLMs too.

The main references

:: Resources (slides, scripts, tasks) and reading

- P. Roback and J. Legler: **Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R.** Chapman and Hall/CRC, 2021.
- P. Rigollet: **Statistics for Applications.** MIT Open Courseware, lecture on GLMs.