B4M36SAN BE4M36SAN Seminar/Tutorial

Alikhan Anuarbekov

Ali

anuarali@fel.cvut.cz

- You have 1 weeks left to form a team (deadline 11.11.2024)
  - 4 people maximum
  - YOU MUST FORM and be preparing for Final assignment
  - I sent you all an email with link to the Team forming google spreadsheet

- Consultation time slots are available on Google spreadsheet as well

- After the deadline you won’t have enough time to discuss your topic!
  - 1 week to come up with an idea, 2 week to prepare a detailed plan

- Also, next weeks you will have a midterm test
  - You can prepare by looking into these files (Courseware – SAN - exam)

5. Solved exercises:
   - san_solved.pdf.
6. Sample questions pertaining to prerequisites:
   - stat_min.pdf, stat_min_eng.pdf.

- **Midterm test**
  - **4 Questions, 60 minute, written on paper during Seminar 8 (11.11.2024)**
  - **1 Question from statistical minimum, 3 from lectures/seminars**
  - **10 points, but if you failed (0 points) = does not matter! Like a big bonus assignment**
    - **You only need 25 points and submit homeworks for non-zero points!**

- **Some questions from previous years:**

1. **What is a p-value?** Problem: it is not a point, but an entire interval area outside of confidence interval
2. **Given data [-1, -2, 0, 1, 3], compute a test against hypothesis = 1.**
   Problem: decide whether to use a t-test or z-test (is variance known?)
3. **Define expected value/arithmetic mean/Bayesian formula…..**
4. **What are Generalized Linear Models? Or Generalized Additive Models?**
5. **Given all estimated coefficients, compute a prediction for a given X1...Xp (maybe not even a simple linear regression, but logistic or something else)**
   Problem: X1, X2 predictors, 3 coefs – 3D graph, do not confuse Y and X2!
6. **Compute log odds from logistic regression formula**
7. **Define spline formula for degree=3 and 2 knot points**

5. Solved exercises:

- san_solved.pdf.

6. Sample questions pertaining to prerequisit

- stat_min.pdf, stat_min_eng.pdf.

A common approach to analyze any linear regression (Recap of assignment 1)

A common approach to analyze any linear regression (Recap of assignment 1)

1) Look to F-stat and R^2/adjusted R^2

```
Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,      Adjusted R-squared:  0.543
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

1.1) If F-statistics has **p-value > 0.05** (lets say ~0.12), then you may actually just stop analysis whatsoever, there is a very high possibility that the model actually is useless since all X predictors have **beta** = 0

1.2) Then you look at **R^2**:

1.2.1) if it is high, lets say **0.85-0.9 (85-90%)**, then the model itself is quite good and maybe you don't even need to modify it

1.2.2) if it is medium, lets say **0.4-0.7 (40-70%)**, it requires a partial modification, remove couple of X predictors and maybe 1-2 nonlinear predictors

1.2.3) if it is low, lets say **0.1-0.3 (10-30%)**, it requires a significant modification and, maybe, most of X predictors are useless or model is highly nonlinear

1.3) Also, **adjusted R^2** tells you actually about predictor impact, if it is decreased significantly in comparison to R^2, then it means that you should do a Feature selection – you have only some significant X predictors

A common approach to analyze any linear regression (Recap of assignment 1)

1) Look to F-stat and R^2/adjusted R^2

```
Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,        Adjusted R-squared:  0.543
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
Residual standard error: 0.7181 on 496 degrees of freedom
Multiple R-squared:  0.7493,  Adjusted R-squared:  0.7488
F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

1.4) If F-statistics of two models have large difference in value (left), but ~ same p-value, THEY ARE ~ SAME
(SAME = same order of magnitude, e.g. **2.2e-16 and 1.2e-16** are SAME, while **2.2e-16 and 1.2e-4** NOT

Or, if you actually want to use them, you could compare these F-statistics using ANOVA,
this method will actually tell you if there is any significant difference or not (you will get combo p-value):

$$ANOVA\left(model_1, model_2\right) = \frac{F-stat_1}{F-stat_2} = \frac{S_1}{S_0} \div \frac{S_2}{S_0} = \frac{S_1}{S_2} = F-stat$$

A common approach to analyze any linear regression (Recap of assignment 1)

2) Look to p-values and number of samples

ical) status in the given town. The model was built from a training set based on 506 towns and is this:

```
lm(formula = medv ~ lstat, data = Boston)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.1) A simple rule: if you have more than 2 variables, do not use p-values as a method of feature selection

1 – (1 – 0.05)^num_variables:

1 variable = 5% type I error
2 variables = 10% type I error
3 variables = 15% type I error
4 variables = 19% type I error

2.2) Also, to use p-values you need N = number of samples (in this case 506) much larger than K = number of predictors (1 in this)

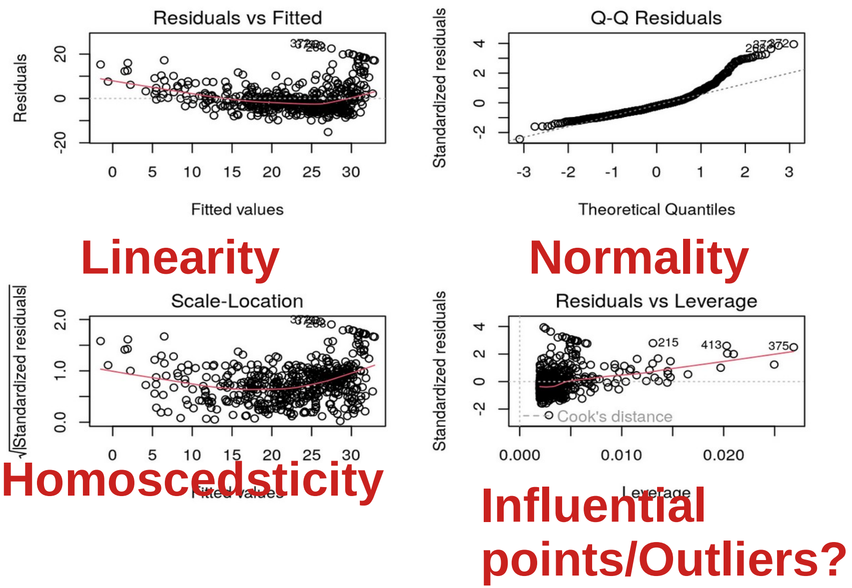
2.3) Otherwise = maybe try p-values along with other Feature selection

A common approach to analyze any linear regression (Recap of assignment 1)

3) Look to plots and tests

3.1) If at least one of them is significantly violated, then you should either:

3.1.1) Try to solve it:
*) **Cook distance ->** remove outlier
*) **VIF** -> find highly correlated pairs, remove by hand
*) **Transformation** (log/square root/BoxCox):
Solve non-linear/ reduce heteroscedasticity
*) Use **weighted least squares** (**WLS**)
(linear regression with in-built heteroscedasticity)



3.1.2) Or if can't solve all, use non-linear models:

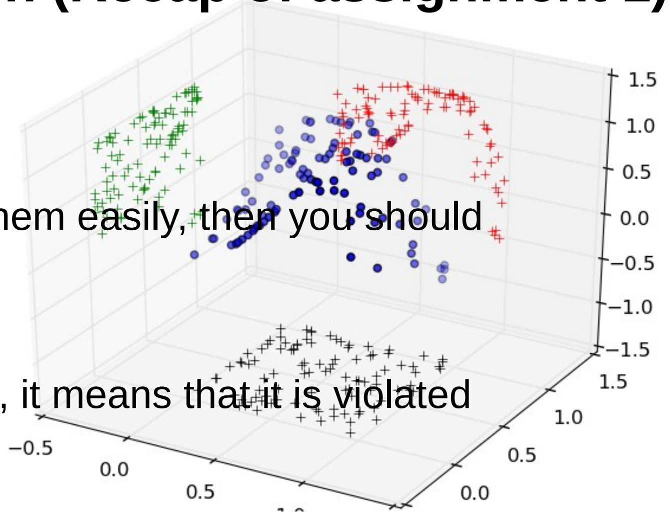
3.1.2.1) **Generalized Additive Models (GAMs)** = polynomials, step functions, splines

3.1.2.2) **Generalized Linear Models (GLMs)** = Poisson/log-Normal/Binomial/Logistic/….

3.1.2.3) Or a combination of **GAMs** + **GLMs**

A common approach to analyze any linear regression (Recap of assignment 1)

4) If too violated, check the problem variable

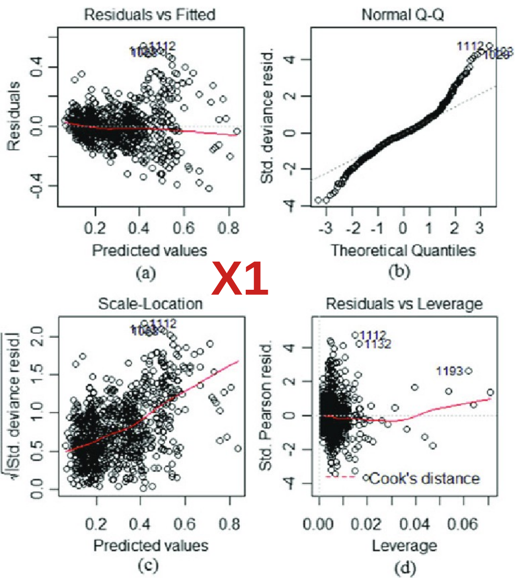


4.1) If you have found that assumptions are violated and you cannot solve them easily, then you should

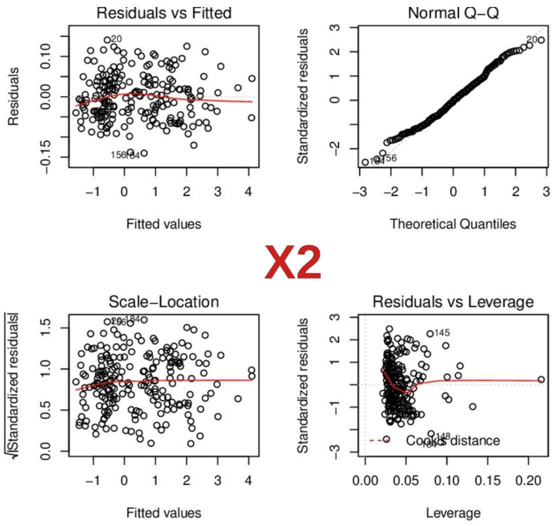
4.1.1) Try to perform same tests, but for each variable independently

*) If it is not violated, just use plain linear model for this variable

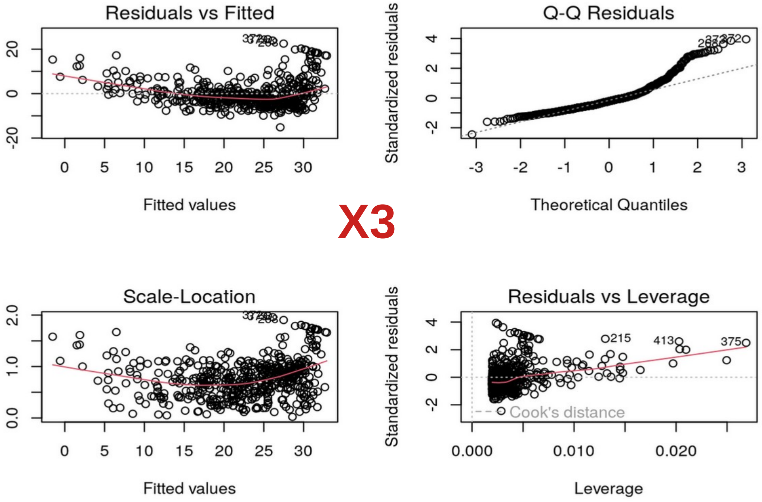
*) Remember the projection principle – If it is violated in residuals, it means that it is violated in some of variables (not necessary all)



X1



X2



X3

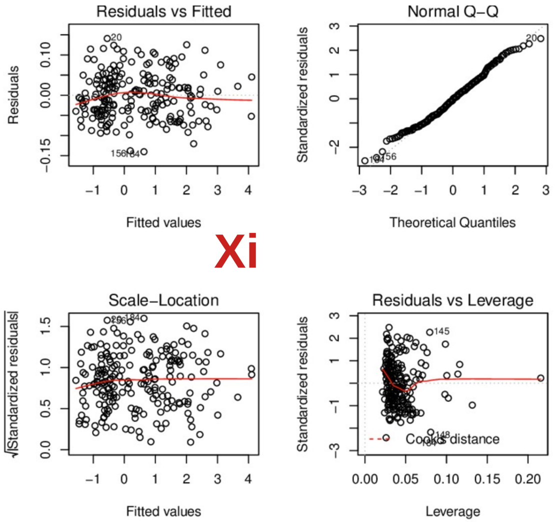
A common approach to analyze any linear regression (Recap of seminar 4)

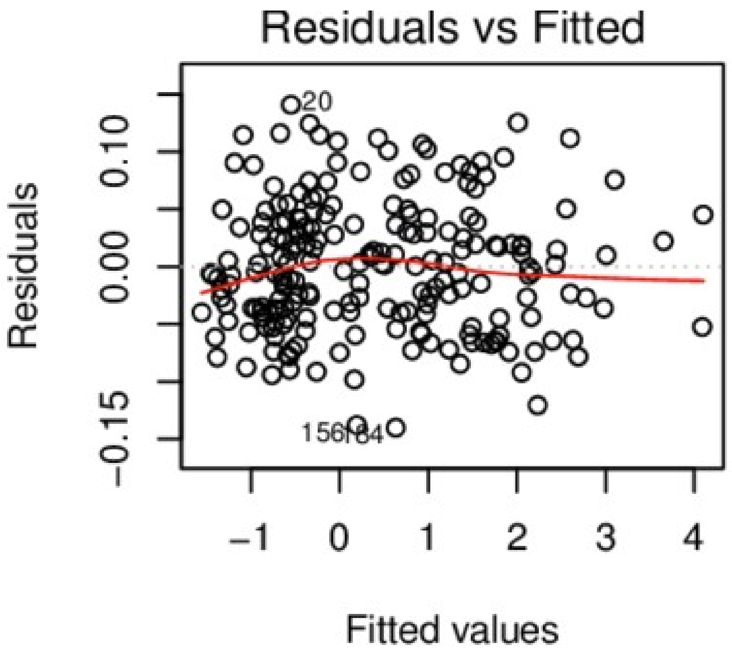
4) If too violated, check the problem variable

4.2) For each variable look into the first plot (or you can just use Xi – Y plot)

 4.2.1) First plot gives you the “smoothing” line of means, predicting the shape of possible polynomial

 4.2.2) We can use it to try to derive the type

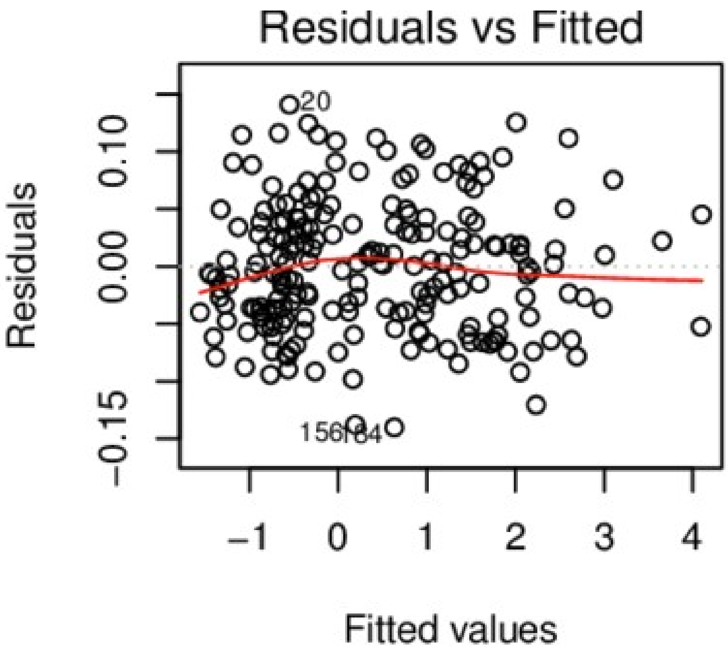


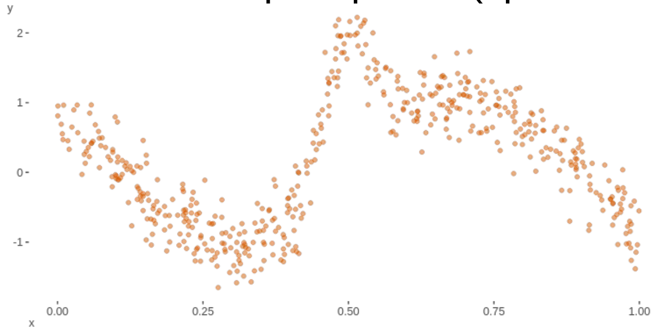


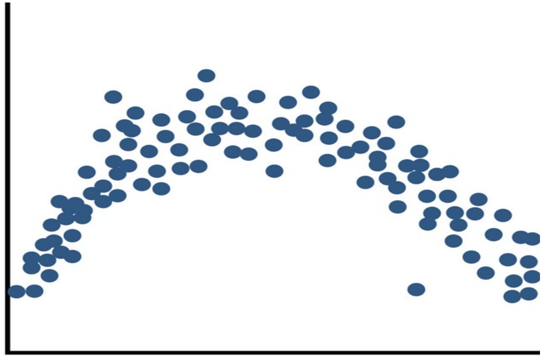
A common approach to use Generalized Additive Model (GAM, nonlinear reg.) (Re

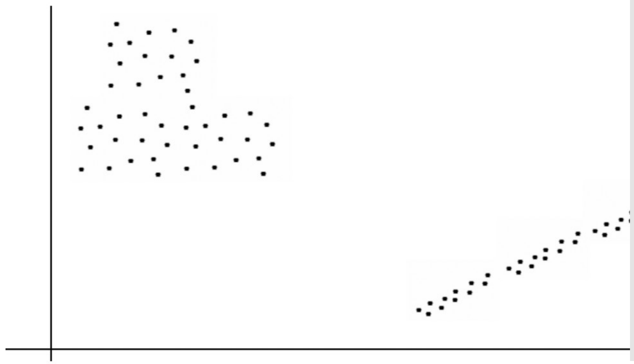
4) If too violated, you can use GAMs

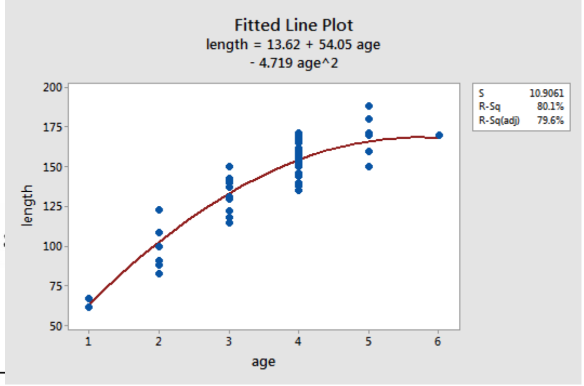
4.2.2) Checking the type: based on shape you can have multiple options (spline/ step/ piecewise poly):





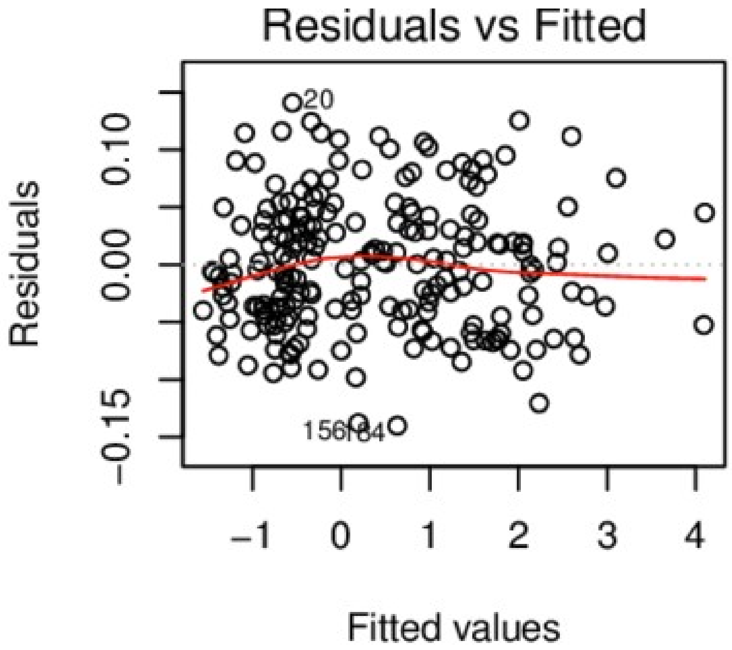






A common approach to use Generalized Additive Model (GAM, nonlinear reg.) (Re

4) If too violated, you can use GAMs



4.2.3) Finding out what is the optimal model among possibilities:

I think that model can be:

1) simply linear with some deviations
2) single polynomial, maybe wide parabola
3) smoothing spline

…..

Option A:

Just try EVERYTHING THAT IS POSSIBLE in a for loop:

models = {poly (2,3,4,5,6,7,…), spline (2,3,4,5,5, knots=2,3,4,5

```
for model in models:
        train(model)
        MSE/AIC/... = test(model)
```

result = best MSE/AIC/…. Model -----------------------> IDEALLY BEST ACCORDING TO MULTIPLE CRITERIA

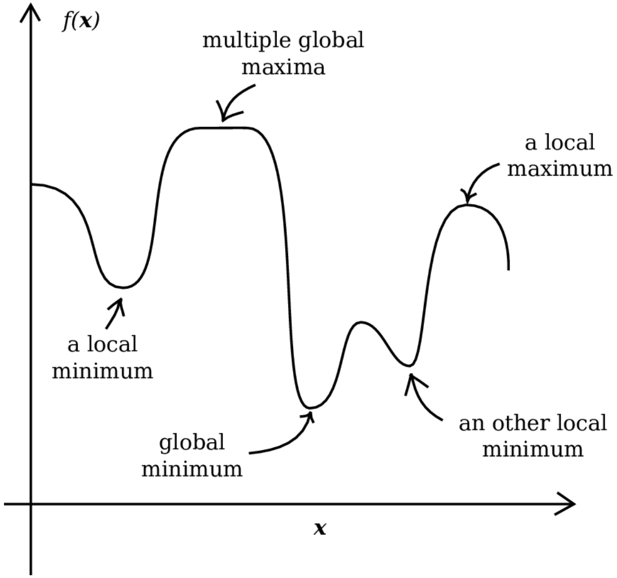
A common approach to use Generalized Additive Model (GAM, nonlinear reg.) (Re

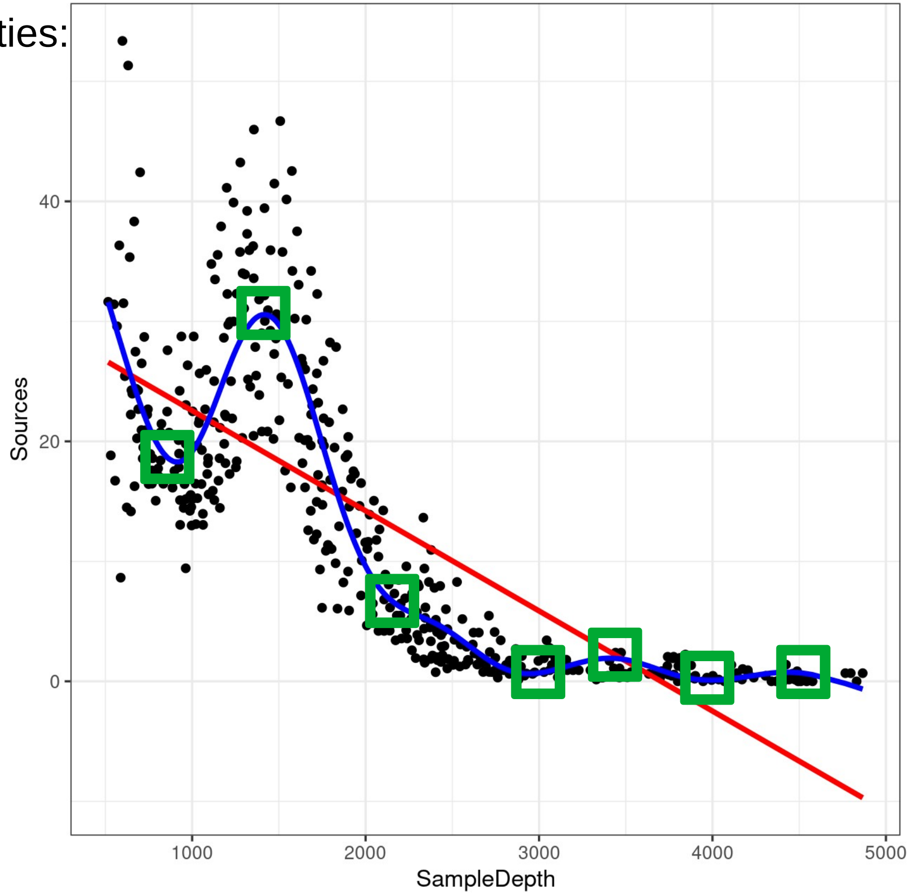
4) If too violated, you can use GAMs

4.2.3) Finding out what is the optimal model among possibilities:

Option B:
find REASONABLE models and test them in a for loop

Count number of ***smooth*** inflex/local optimum points
(degree of polynomial):





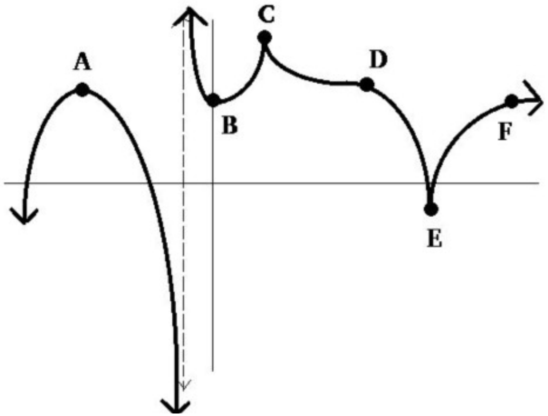
A common approach to use Generalized Additive Model (GAM, nonlinear reg.) (Re

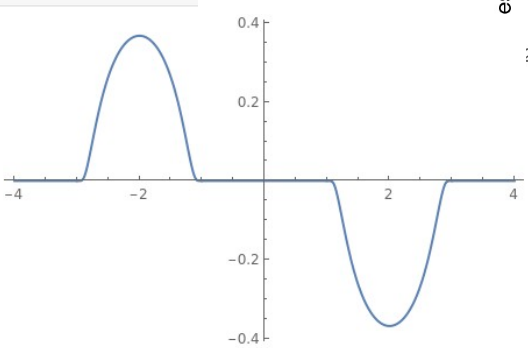
4) If too violated, you can use GAMs

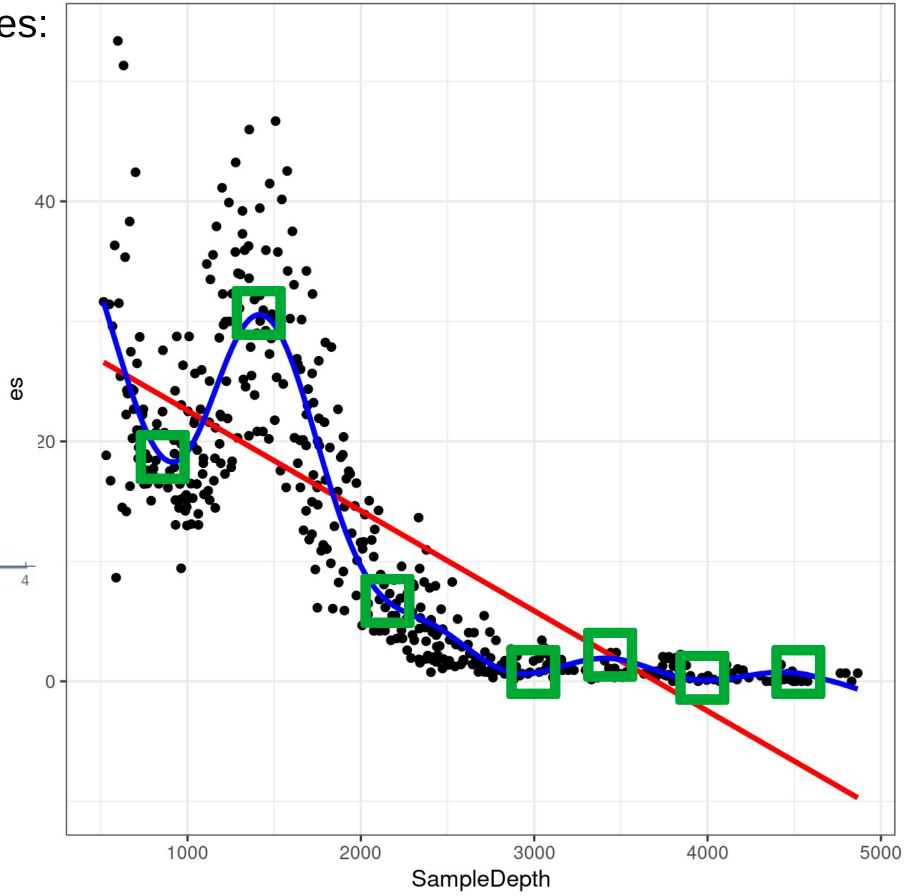
4.2.3) Finding out what is the optimal model among possibilities:

Option B:
find REASONABLE models and test them in a for loop

But also number of ***not-so-smooth*** points or jumps
(this can be knots of spline or step jumps):







A common approach to use Generalized Additive Model (GAM, nonlinear reg.) (Re

4) If too violated, you can use GAMs

4.2.3) Finding out what is the optimal model among possibilities:

Option B:
    find REASONABLE models and test them in a for loop

Given that you compute number of ***<u>smooth optima/inflex points</u>*** and ***<u>not-so-smooth points</u>***,

Try all of previous model types: linear, step, polynomial, spline, ….

But with **degree** and **knots** adjusted according to ***<u>smooth/optima/inflex</u>*** and ***<u>not-so-smooth points</u>***

For example:
    Try one single polynomial with degree = smooth optima/inflex points

    Divide smooth points to intervals of 2-3, e.g. for lets say 12 smooth points use 4 polynomials of degree 3

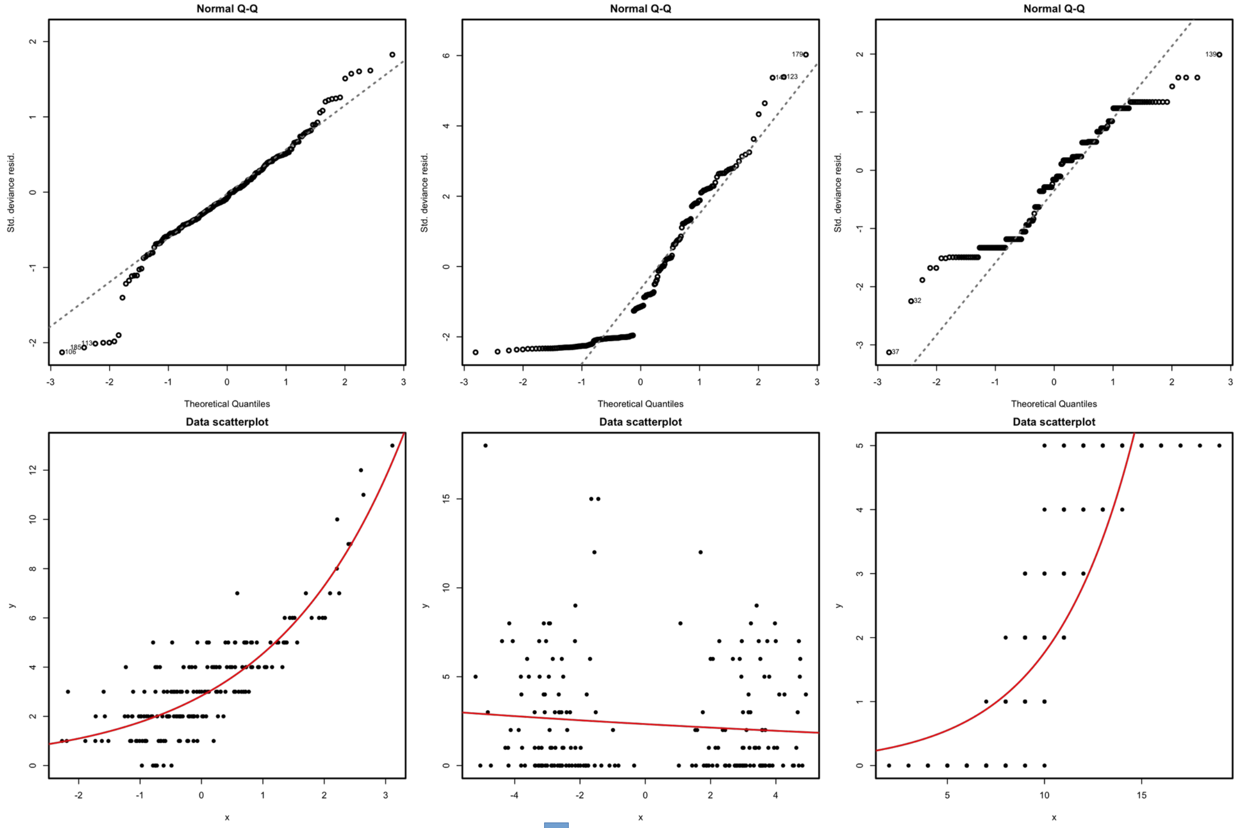
A common approach to use Generalized Linear Models (GLMs)

5) If too violated, you can use GLMs

BUT WHAT IS A GLM?

H = Height, V = age,
D = Number of kittens (babies)

D ~ H + V



Distribution of residuals isn't normal,
sometimes even polynomial transformations wont help (skewed shape because of errors)
maybe try something else than Gaussian normal?

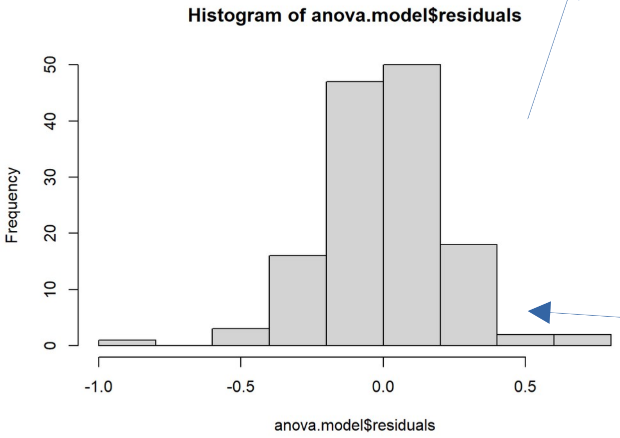
Exploratory data analysis (EDA) for a single X predictor

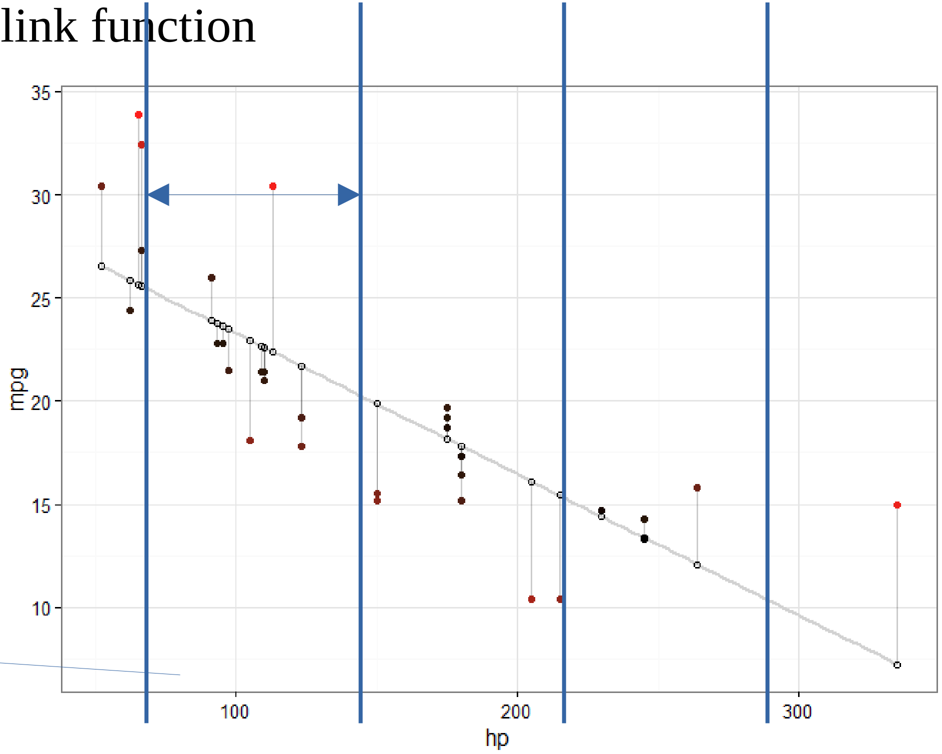
A way of determining the distribution type of residuals $\hat{Y} - Y$

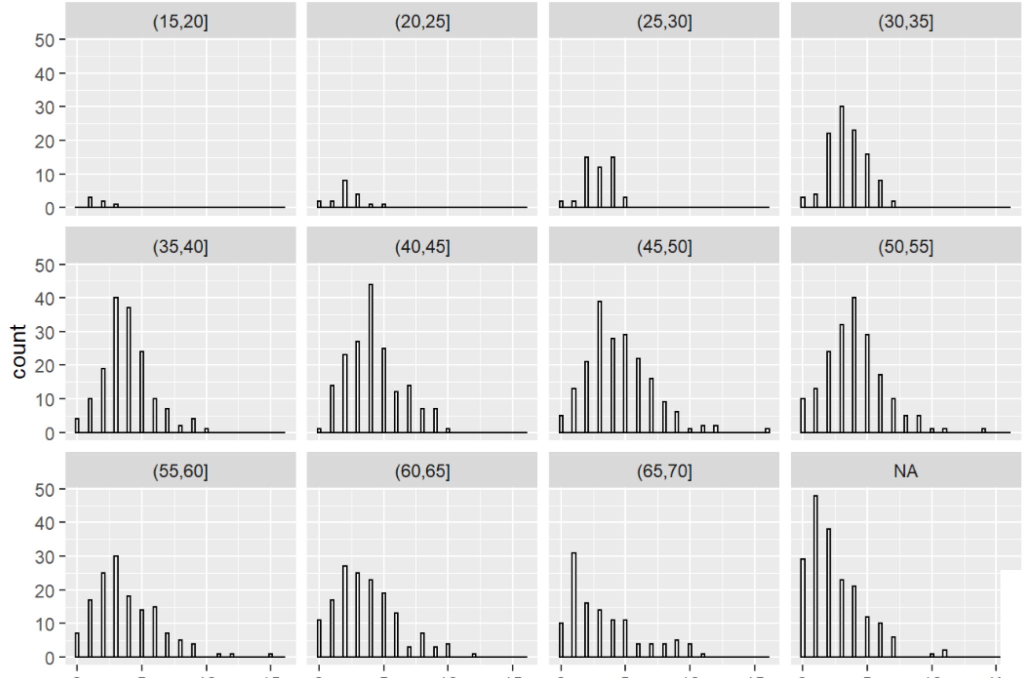
1) Take small interval on X axis such that your have multiple points
2) Compute a histogram of $(\hat{Y} - Y)$ values inside of interval
3) According to histogram shape and properites determine a distribution family
4) According to mean value curve I determine the link function

> intervals because not enough points per each X

Determine distribution type

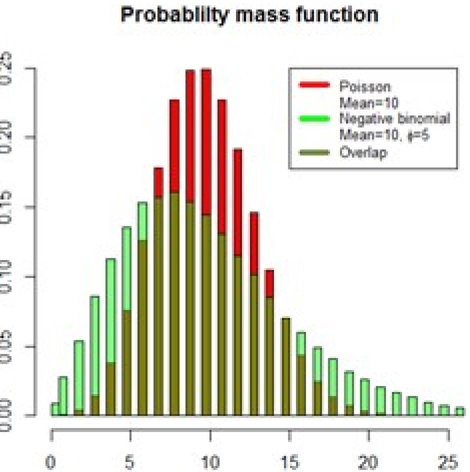


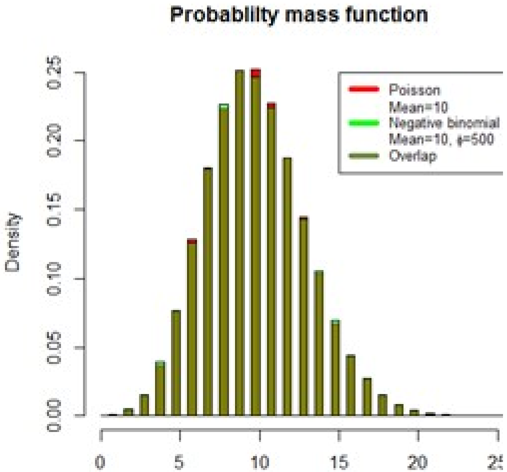


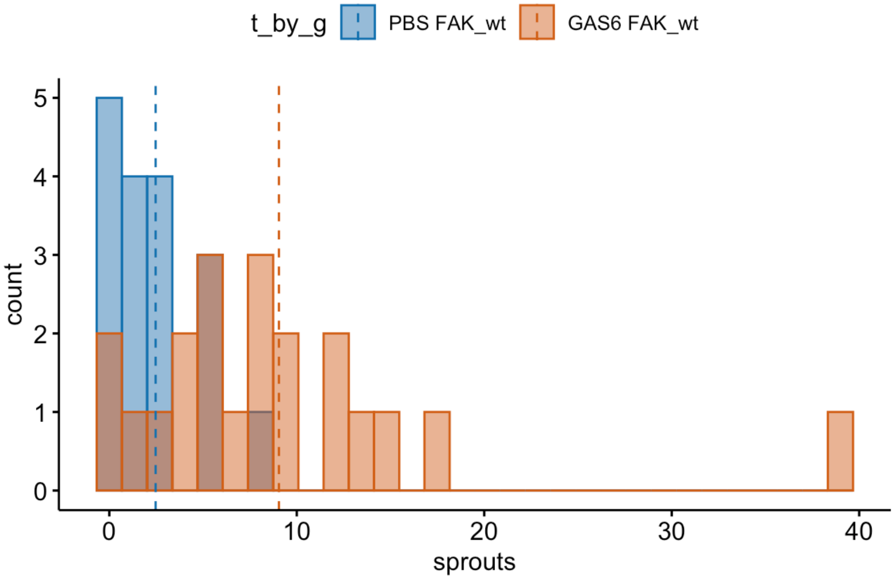


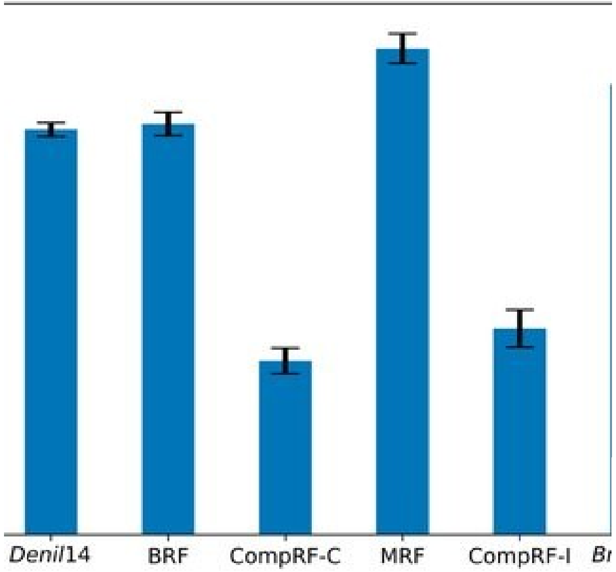
Positive integer values:

Poisson nebo *Negative Binomial*

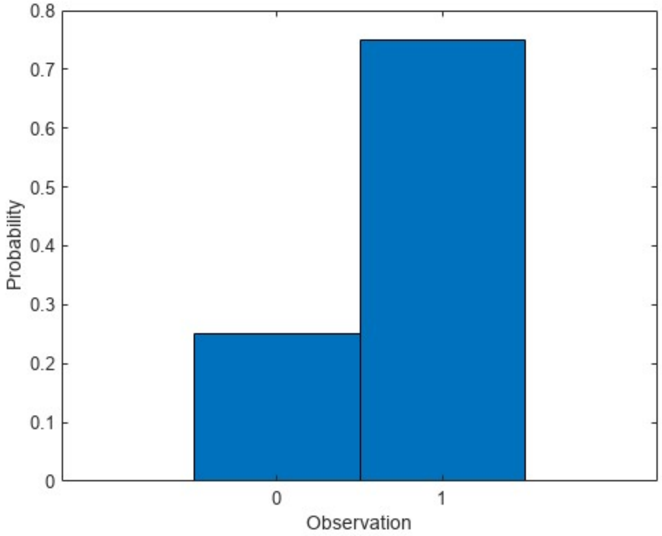


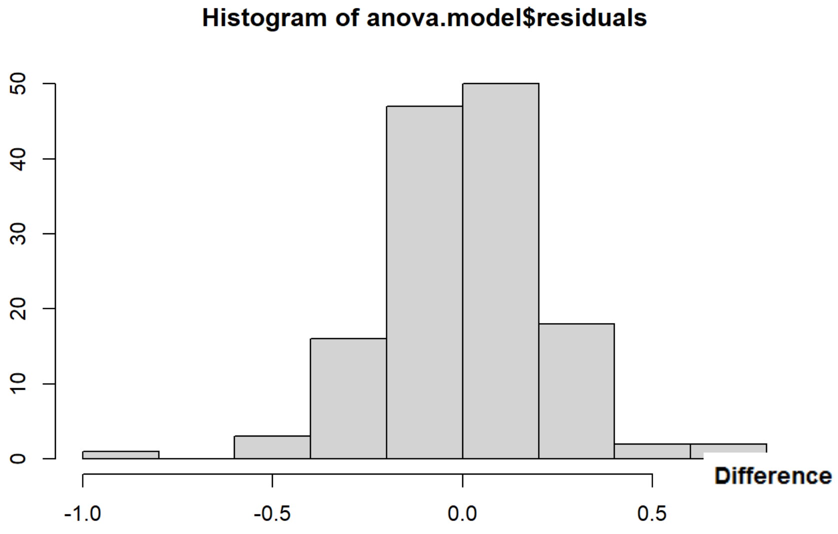






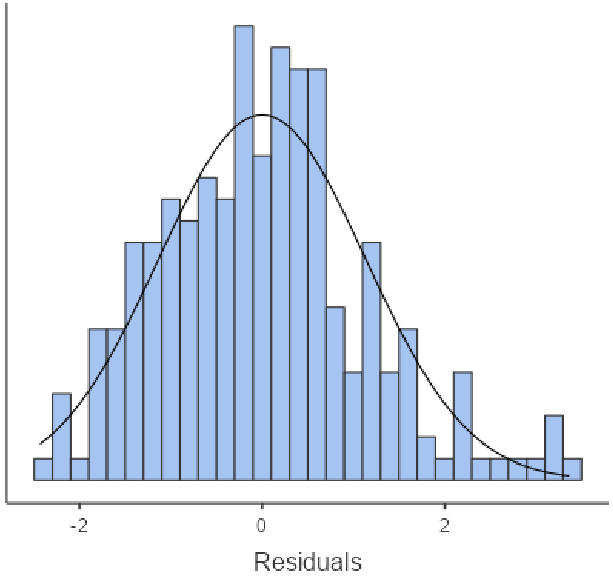
Any integer values with limited values:
$Logistic\,(2\,values)$ nebo $Multinomial\,(\text{more values})$

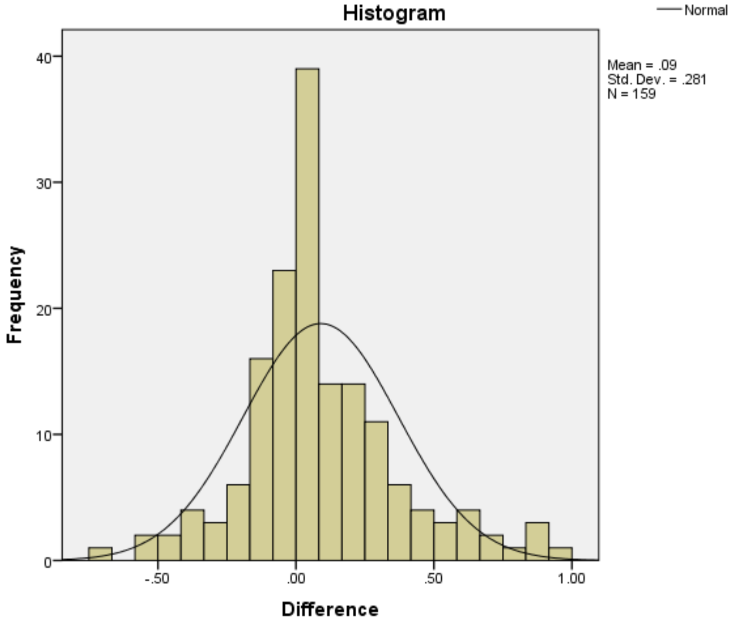


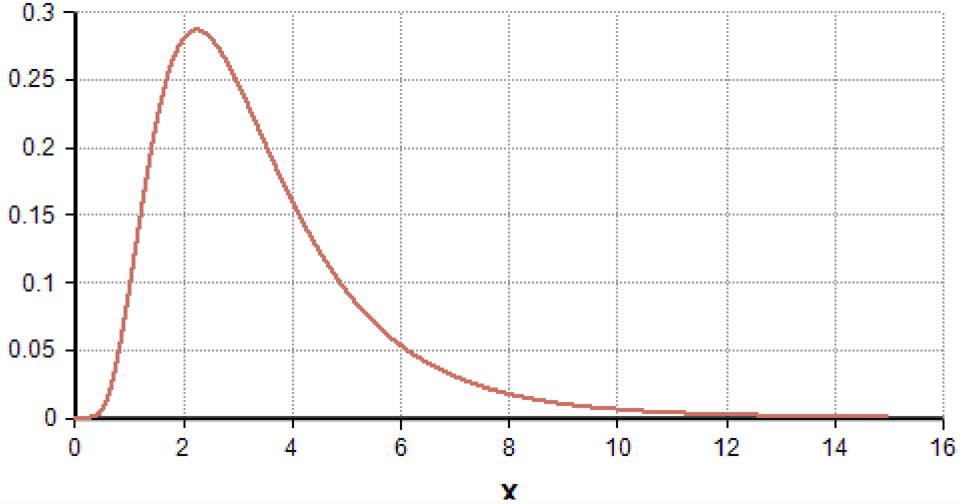


Continuous numbers that are symmetric

Gaussian Normal

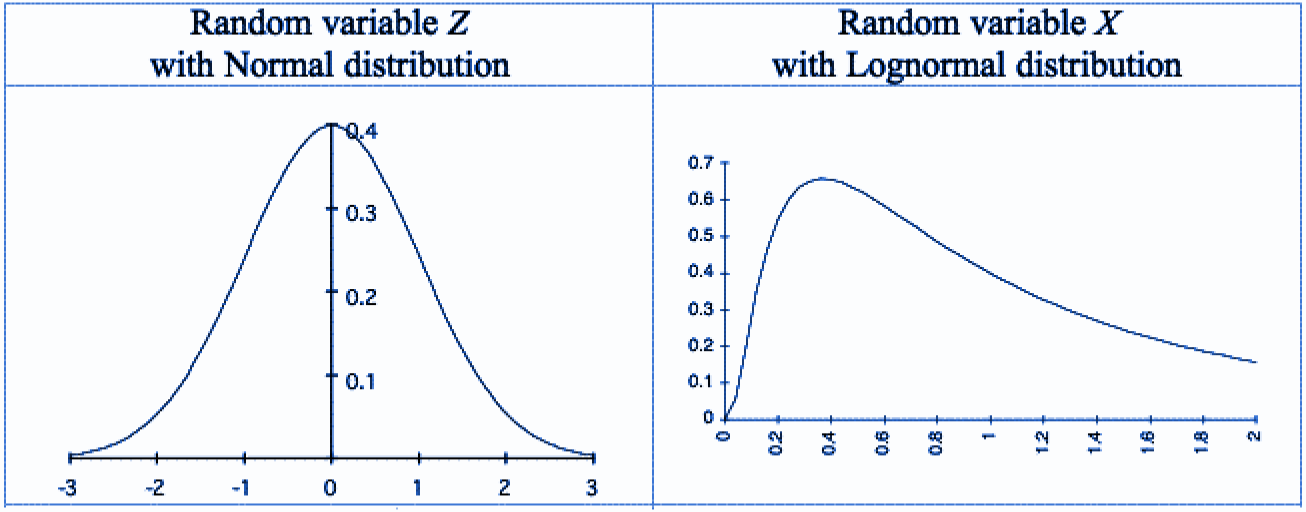




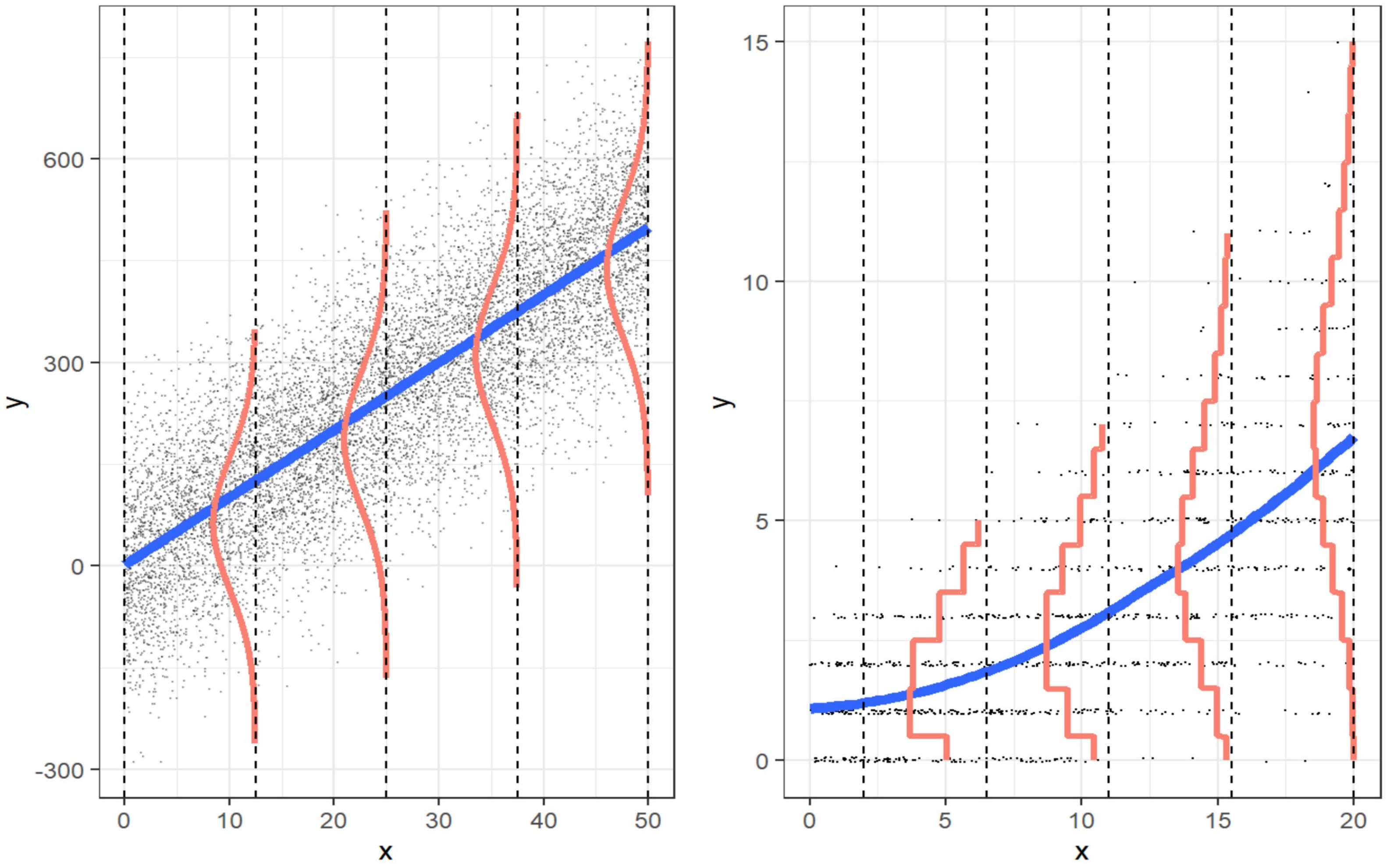


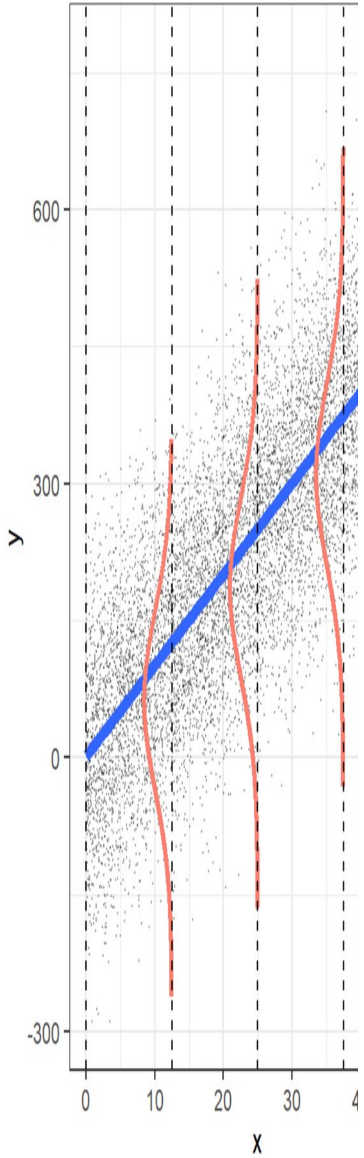
Continuous numbers that are skewed (deformed)

Gamma family = log Normal distribution type



Lets show the full approach on example from lectures:

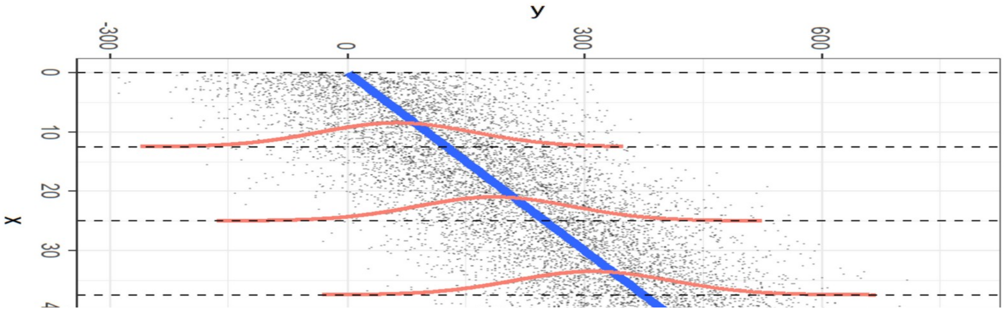


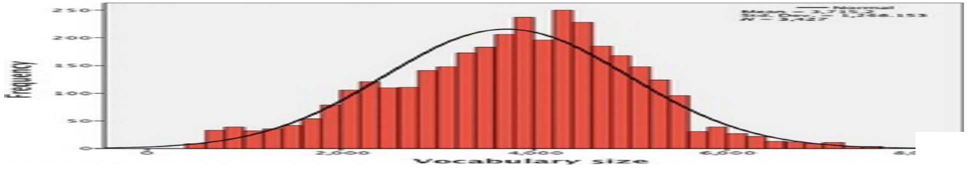


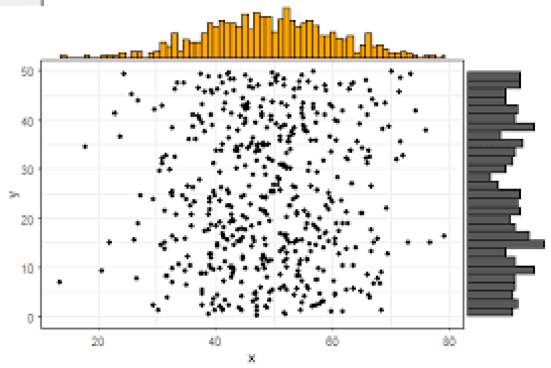
1) Cut some small interval, but such that you have enough points

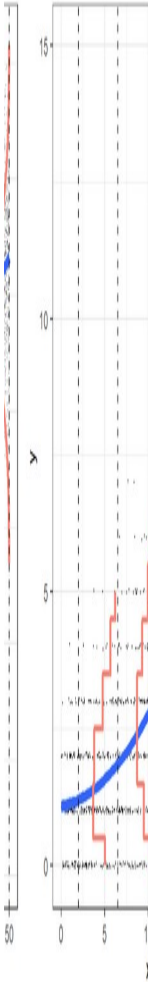
2) Rotate it such that it is more convenient

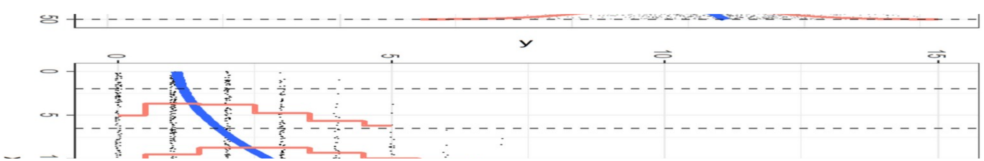
3) Calculate histogram on its projection (count each unique Y value)

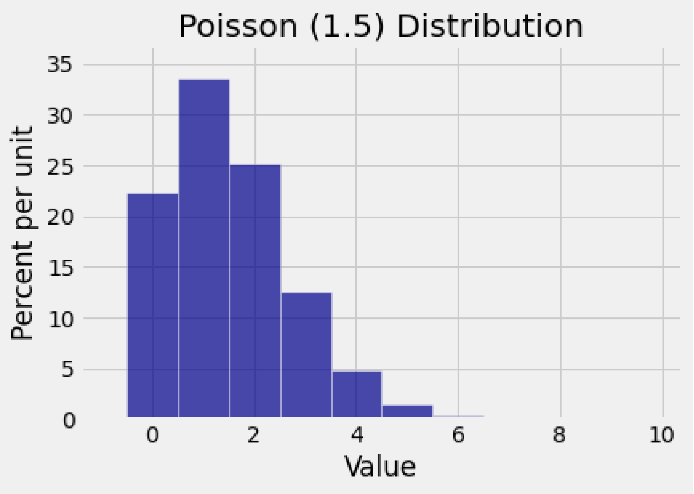




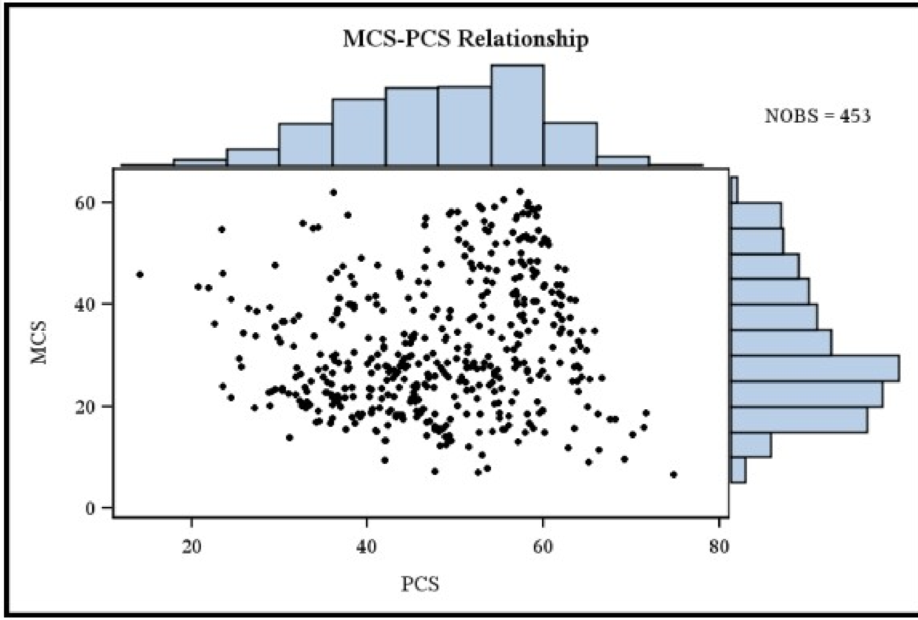






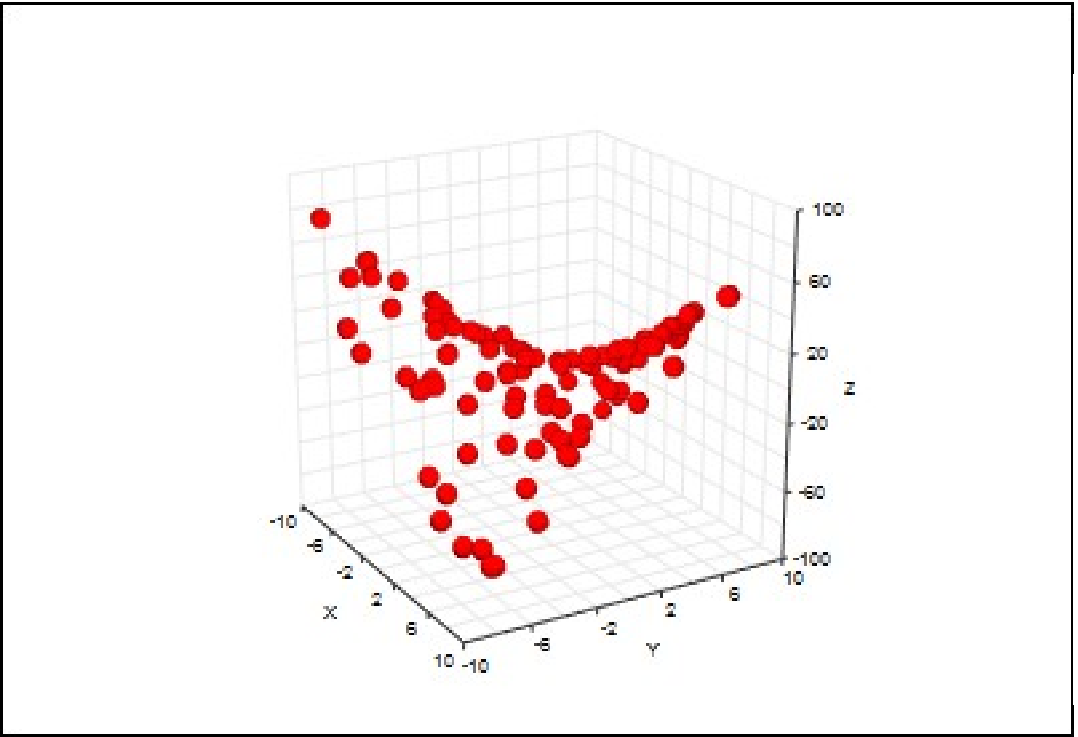


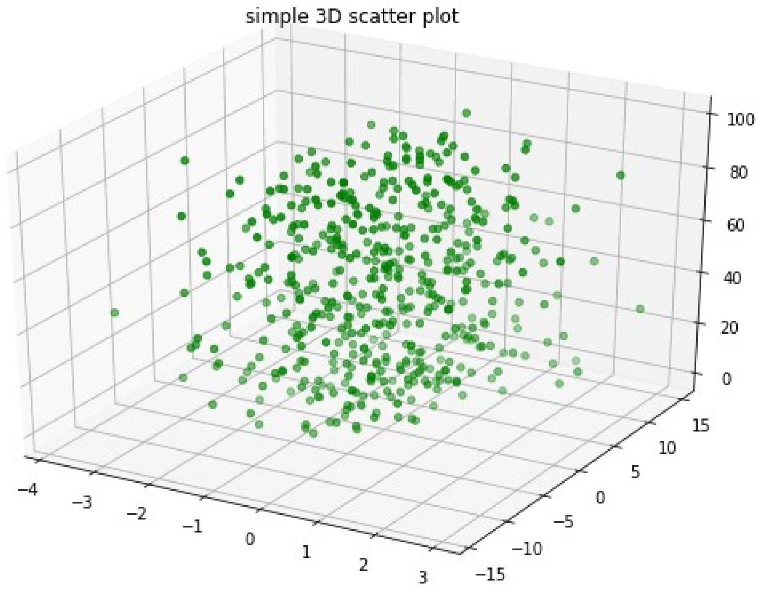
SAME BUT FOR POISSON REGRESSION
(right image from lecture)



Exploratory data analysis (EDA) for multiple X predictors (2 in this case)

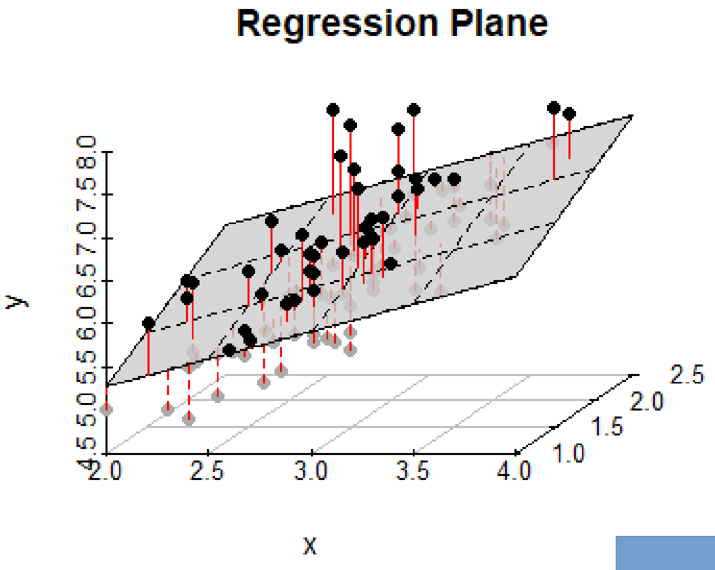
This time we do the same with residuals = $\hat{Y} - Y$



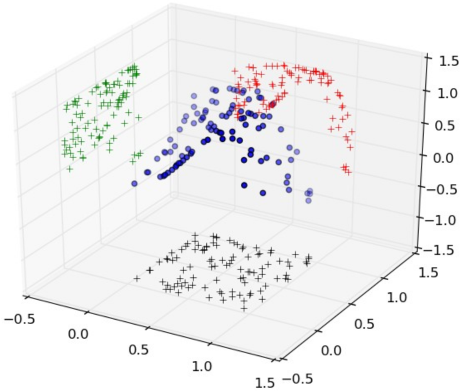


Exploratory data analysis (EDA) for multiple X predictors (2 in this case)

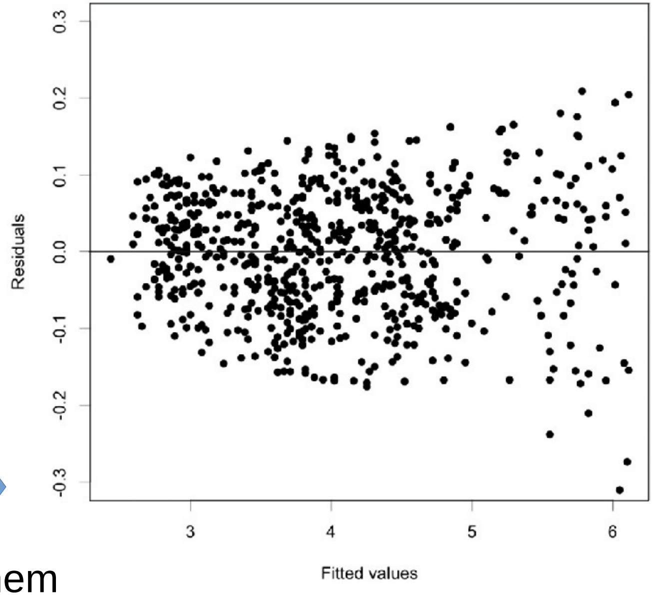
Since we have 2 X axes, we need to do NOT INTERVALS, but GRID $\hat{Y} - Y$



Start with 2D+1D
(or generally N-th dimension)



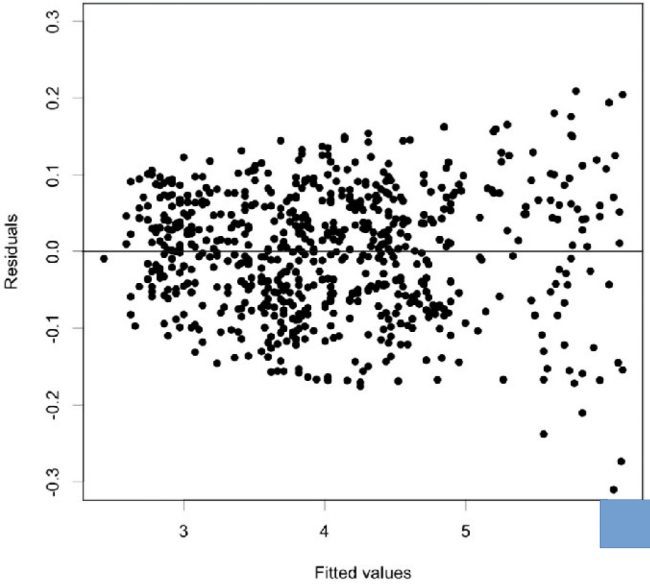
Compute residuals and project them
(For each grid (cell) separately)

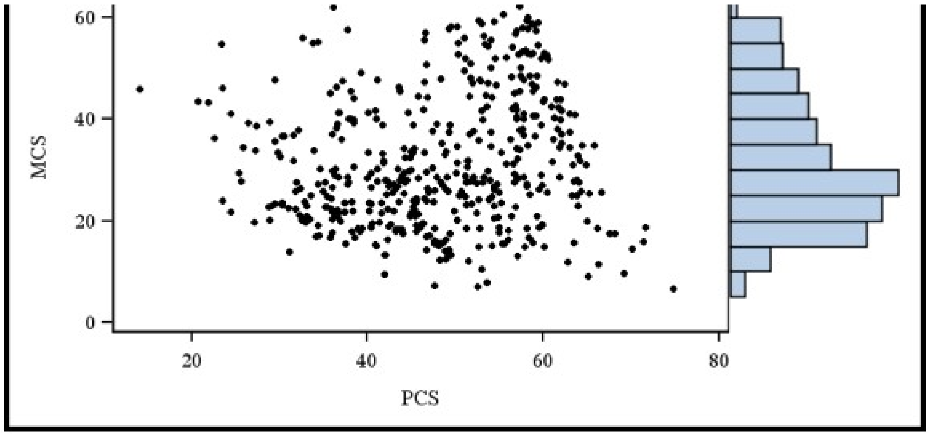


Get a grid residual plot

Exploratory data analysis (EDA) for multiple X predictors (2 in this case)

Since we have 2 X axes, we need to do NOT INTERVALS, but GRID $\hat{Y} - Y$

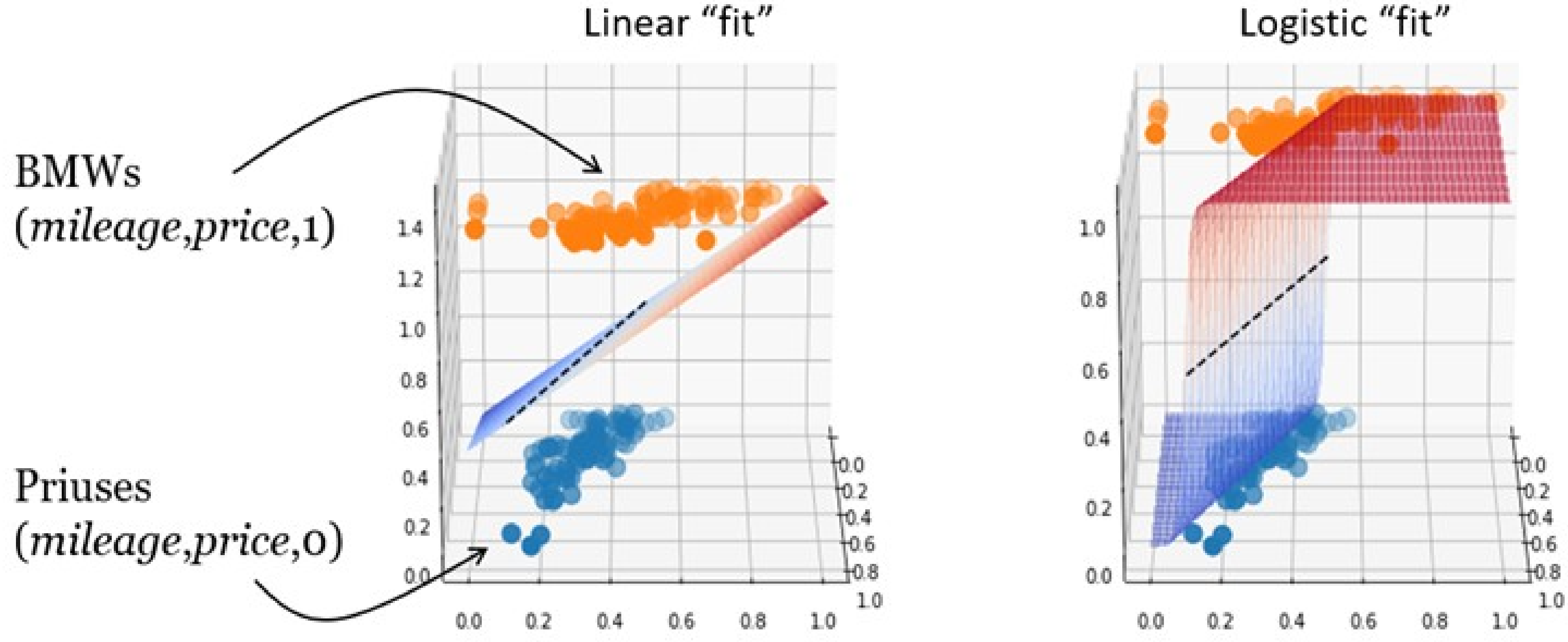




Compute histogram of residuals -> derive distribution type

Get a grid residual plot

Logistic regression is not only in 1D (WAS ON MIDTERM IN PREVIOUS YEARS):



> It can a combination of multiple X: continuous/categorical, does not matter

$$\hat{y}_{lm} = \beta_0 + \beta_1 age + \beta_2 \cdot \text{is\_male} + \beta_3 \cdot \text{insurance (-1, 0, 1)} \rightarrow y_{response} = \frac{1}{1+e^{-\hat{y}_{lm}}}$$

> As long as Y axis is categorical and binary, we can just do LM and apply sigmoid
> Same multivariate approach for any GLM regression!

Let us generalize:

- Recall that with linear regression

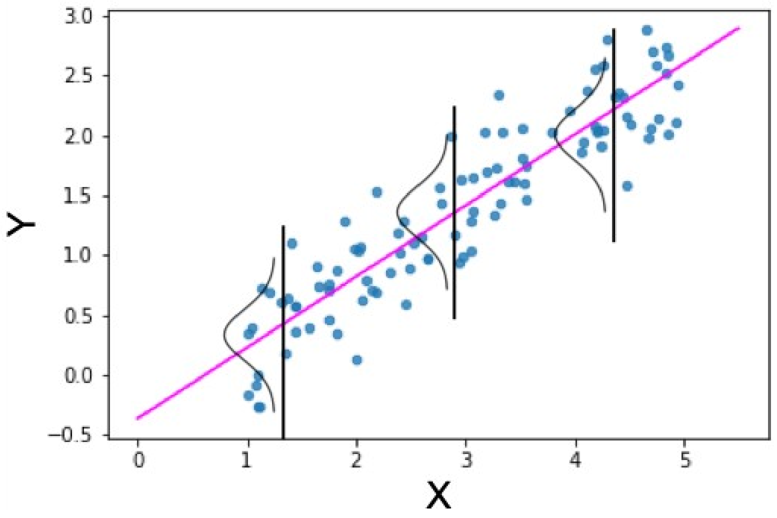
$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$
$$\underline{y_i \sim \mathcal{N}(\mu_i, \epsilon)}$$

- in Poisson regression

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

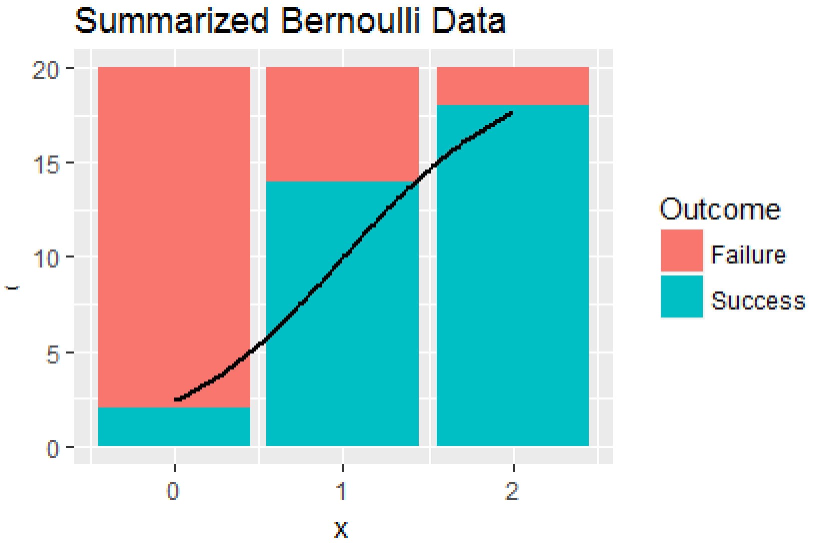
- logistic regression has a similar form

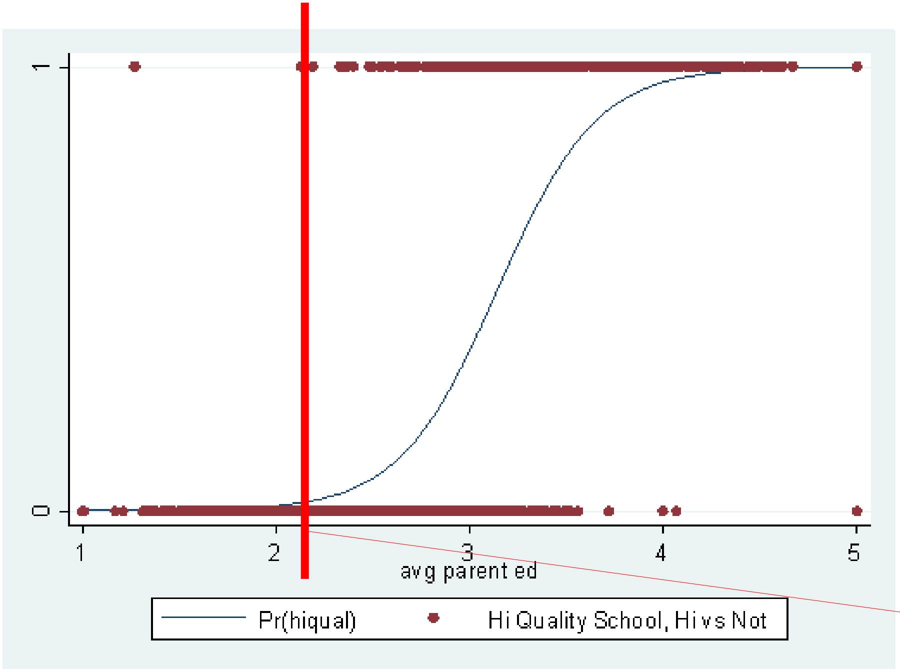
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$
$$q_i = \frac{1}{1 + e^{-\eta_i}}$$
$$\underline{y_i \sim Bernoulli(q_i)}$$



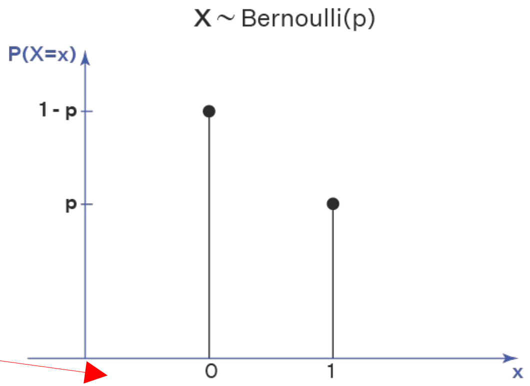
$$f(x)=\begin{cases} p^x*(1-p)^{1-x} & if\ x=0,1 \\ 0 & otherwise \end{cases} = \begin{cases} p & if\ x=1 \\ 1-p & if\ x=0 \end{cases}$$

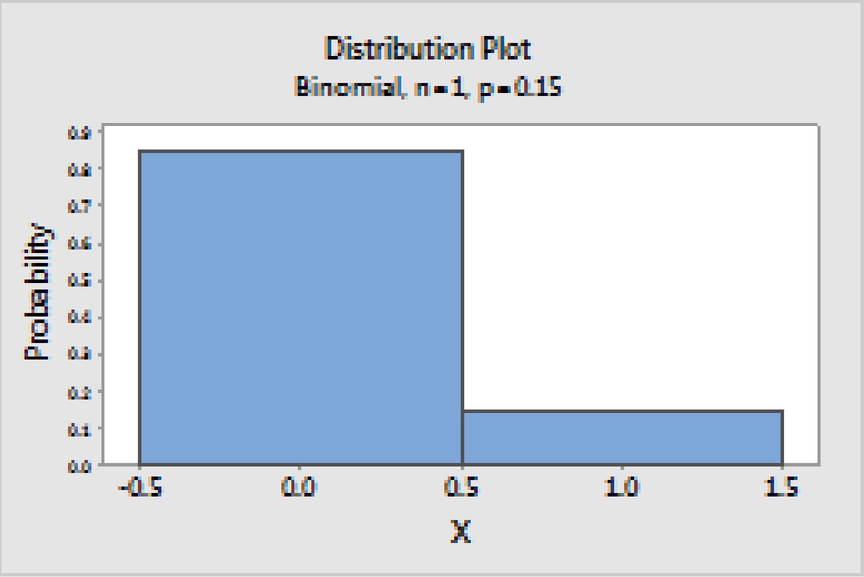
$$K \in [\,0,1\,]$$

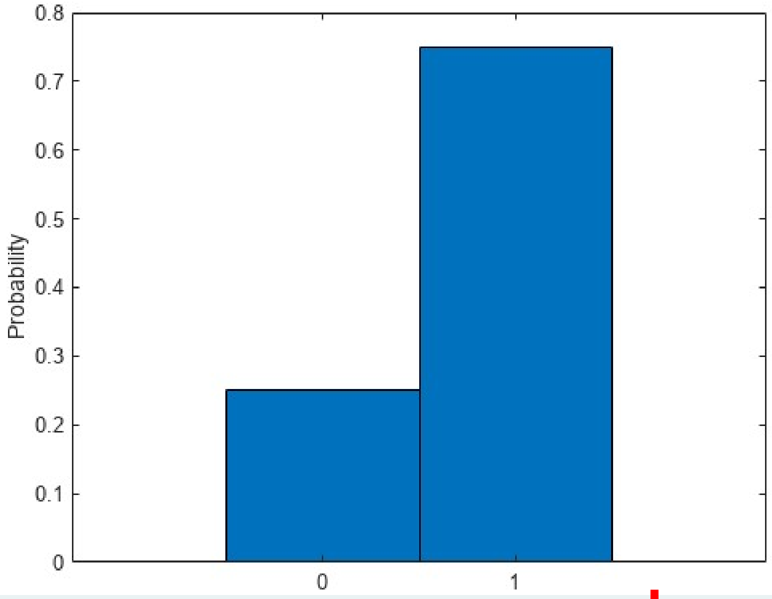


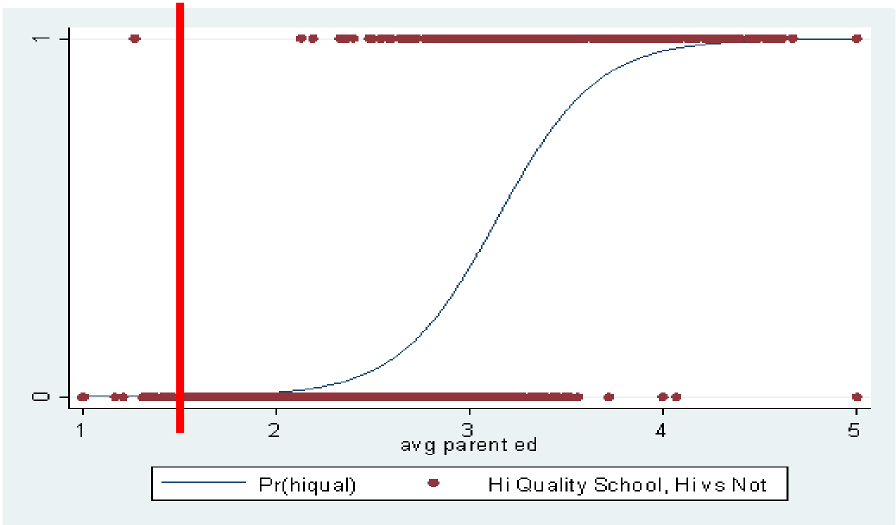


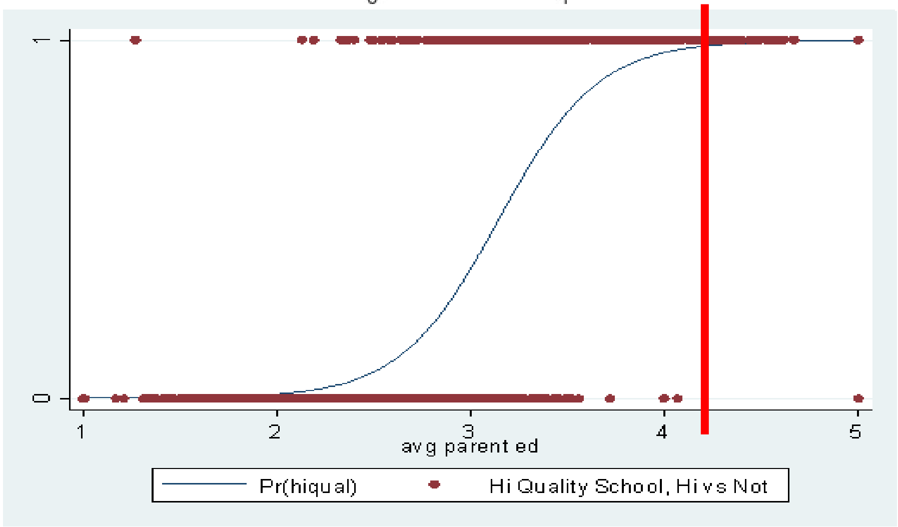
Bernoulli Distribution Graph




Distribution Plot
Binomial, n=1, p=0.15
Probability
X
-0.5
0.0
0.5
1.0
1.5


Probability
0
1


avg parent ed
Pr(hiqual)
Hi Quality School, Hi vs Not


avg parent ed
Pr(hiqual)
Hi Quality School, Hi vs Not

GLMs allow for a more direct transformation functions than polynomials
apply EXP, SQRT, LOGIT or some other transformation, not just “trick of polynomial”

glm options

Family	Link Function	When to use it:
• gaussian (link = "**identity**")	μ	homogeneous and normal residuals, real numbers, may be negative
• poisson (link = "**log**" *or* "identity" *or* "sqrt")	ln(μ)	variance ~ mean, non-normal residuals, count data (>0, integers) (some packages also allow related negative binomial)
• Gamma (link = "**inverse**" *or* "identity" *or* "log")	μ^{-1}	variance ~ mean, even more skew in residuals, real numbers (>0)
• inverse.gaussian (link = "**inverse**" *or* "identity" *or* "log")	μ^{-2}	variance ~ mean, even more skew in residuals, real numbers (>0)
• binomial (link = "**logit**")	ln(p/(1-p))	binary response predicted by quantitative variables; logistic regression
• quasi, quasibinomial, quasipoisson	several	can address overdispersion not in above

6) Use a combination of Generalized Linear Models (GLMs) and GAMs together:

Combine Exponential transformation (log link) + polynomial GAM

```
summary(glm(y ~ poly(x, 6), family = poisson, data = df))

Call:
glm(formula = y ~ poly(x, 6), family = poisson, data = df)

Deviance Residuals:
```

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))
```

Assignment 3: Generalized Linear Models

7	4.11.	JB, AA, JK	Generalized linear models	san_glms.zip, assignment3.zip

Introduction

The aim of this assignment is to practice constructing linear models. You will start with a simple linear model. You will evaluate and interpret it (1p). Consequently, your task will be to improve this model using generalized linear models (GLMs) and feature transformations. You will get 1p for proposal and evaluation of GLM (family, evaluation, interpretation), 1p for correct feature transformations, 1p for proposal and justification of the final model and eventually, 1p for comprehensive evaluation of all the model improvements (ablation study through cross-validation, note that the previous evaluations must be done without cross-validation).

To summarize:

Same as assignment 1, but this time you will use GLMs, GAMs and transformations

Seminar 7: glms.html

Further questions:

1. Show how to refine the model and increase its performance even further. Hint: You may consider other non-linear forms, different link functions and regression types as well as the region feature that has been kept away until now.
2. Compare the models you considered with different criteria and understand how far they match or disagree. Finally, recommend the best model and write a justification for this recommendation.

Small help:

- Just perform the same Exploratory Analysis and choose appropriate GLM family + link + GAM
- Then compare models using AIC/Cross-validation MSE
- Send it to my **Email**