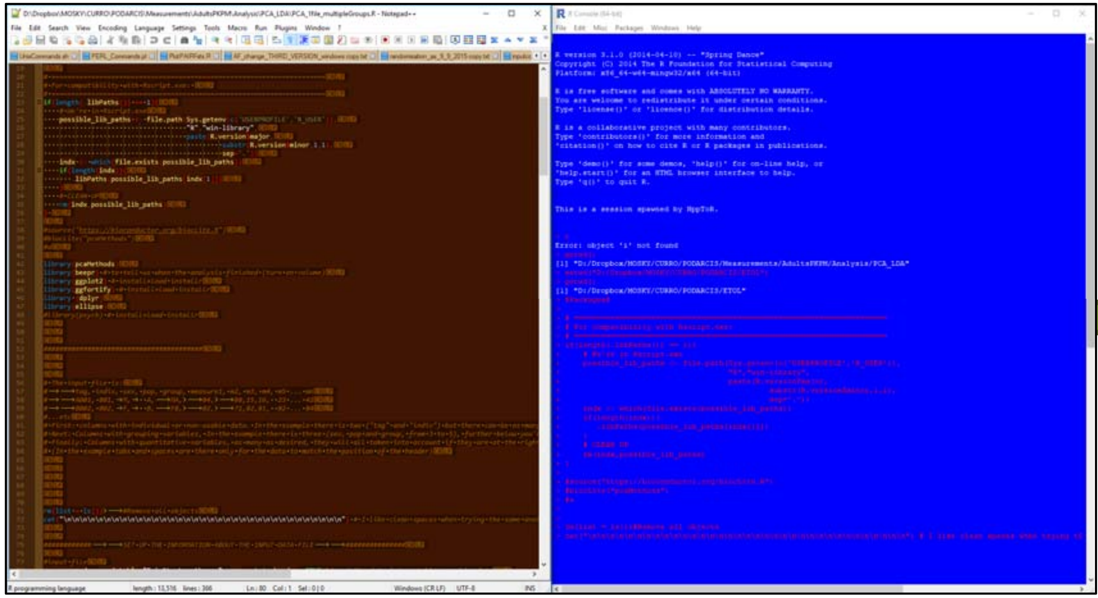
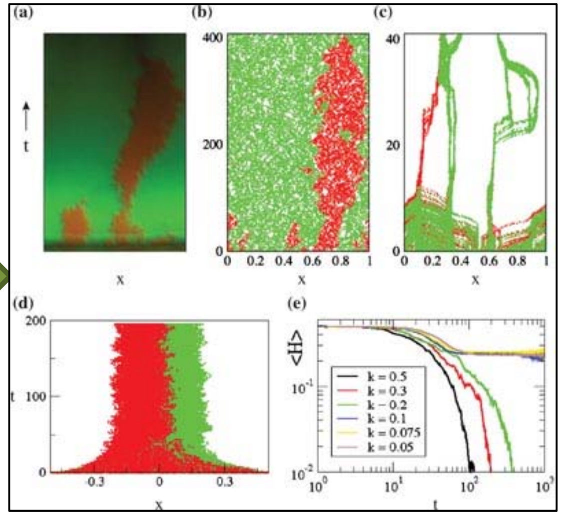
Use of R environment in
Evolutionary Ecology





Remember this?

<u>ANOVA</u>

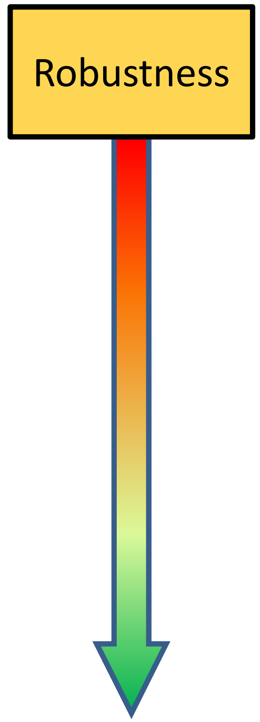
Continuous (numerical)

Samples are chosen randomly

Normal distribution for each group

Homoscedasticity

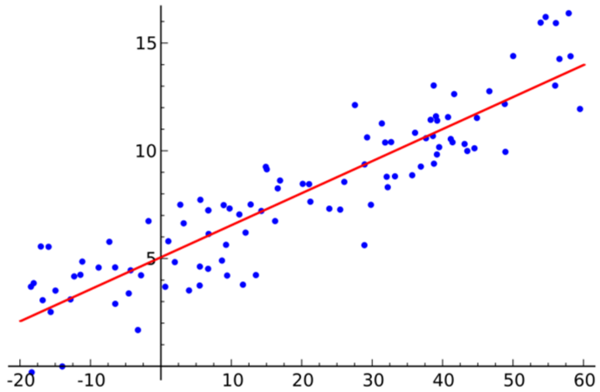
Degrees of freedom = $n-1 \geq 2$



- Preferred because more powerful finding differences
- Data needs to be Normal to fit the predicted distribution
- One of many analysis based on the general linear model

GENERAL LINEAR MODELS

Simple analysis based on linear regression



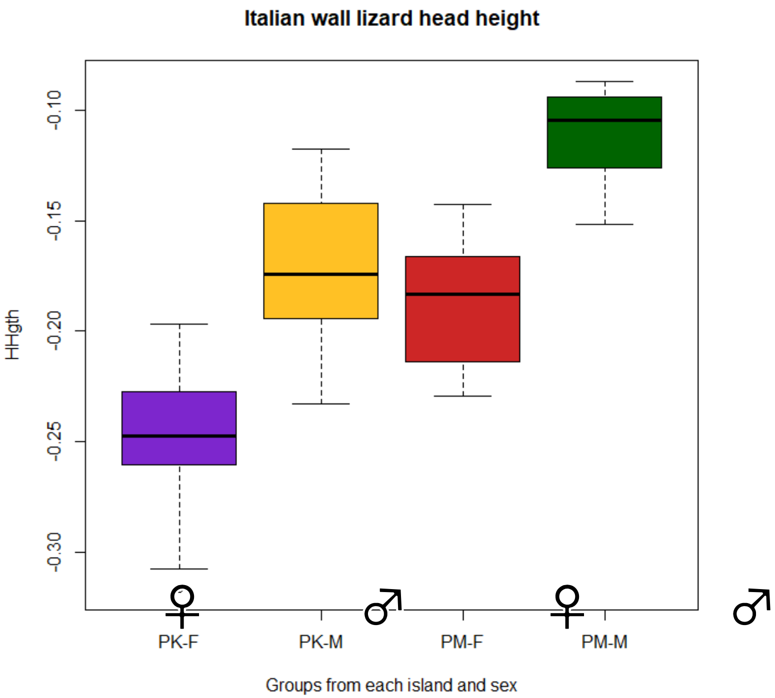
- can **not** handle **not** continuous data
- can **not** handle **not** Normal data
- Can only handle a few variables at the same time

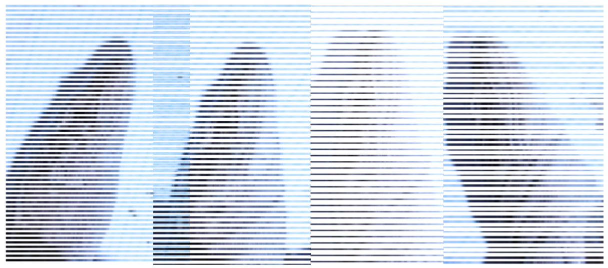
ANOVA, ANCOVA, MANOVA, MANCOVA, t-test, F-test...

MIND SETUP FOR LINEAR MODELS

Categorical variable = **explanatory variable**: Group (Sex+Population)
Continuous variable = **response variable**: Head Height

Are there significant differences in the response variable between groups?





MIND SETUP FOR LINEAR MODELS

Until now the question was:

Are there significant differences in the response variable (Head Height) between groups?

How linear models work:

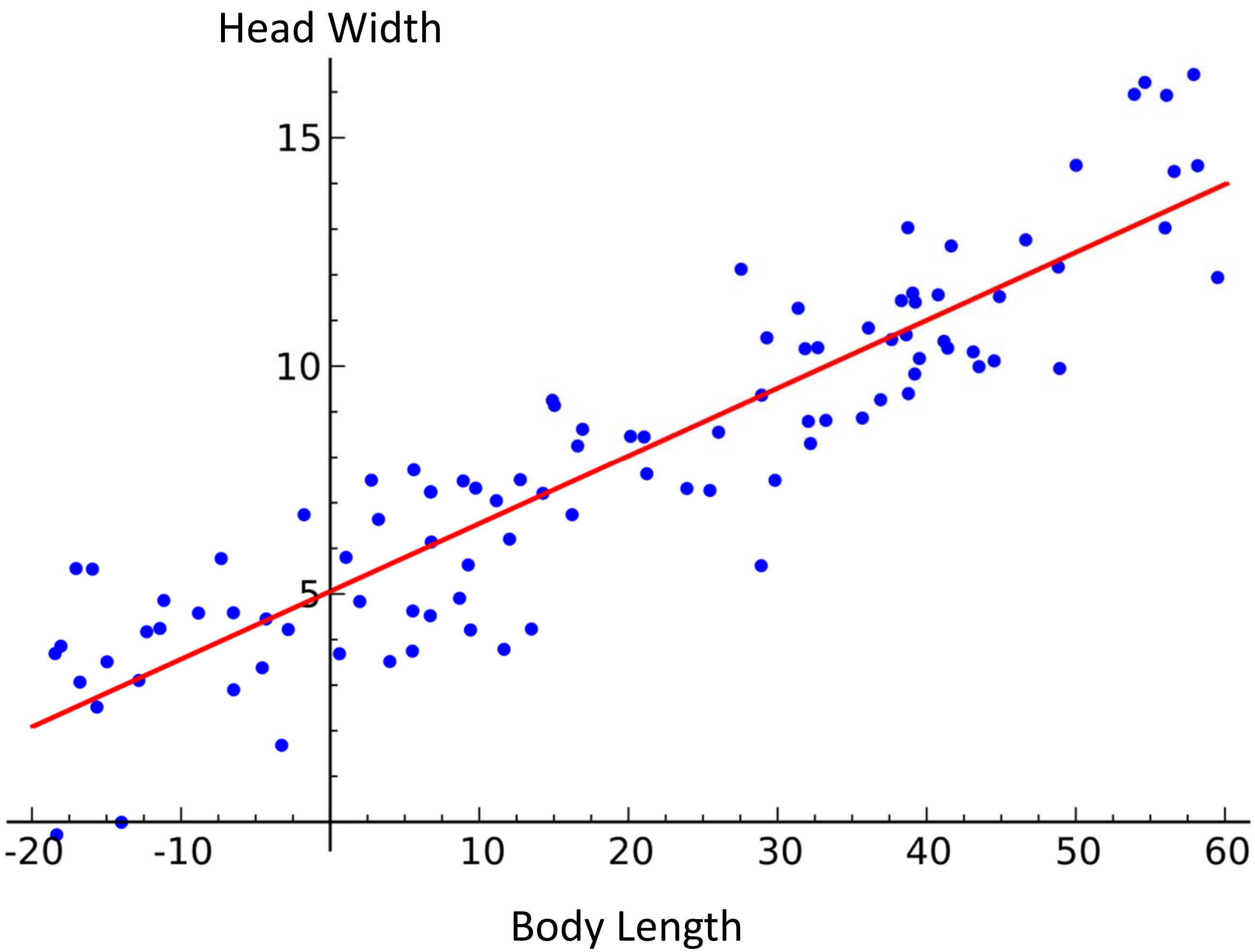
1.- Our data is adjusted to a linear model **y = ax + b**

2.- The model **PREDICTS** the expected values of Head Width acording to another variable (Group, Body Length, etc.)

3.- Then **compares expected** values with real values

MIND SETUP FOR LINEAR MODELS

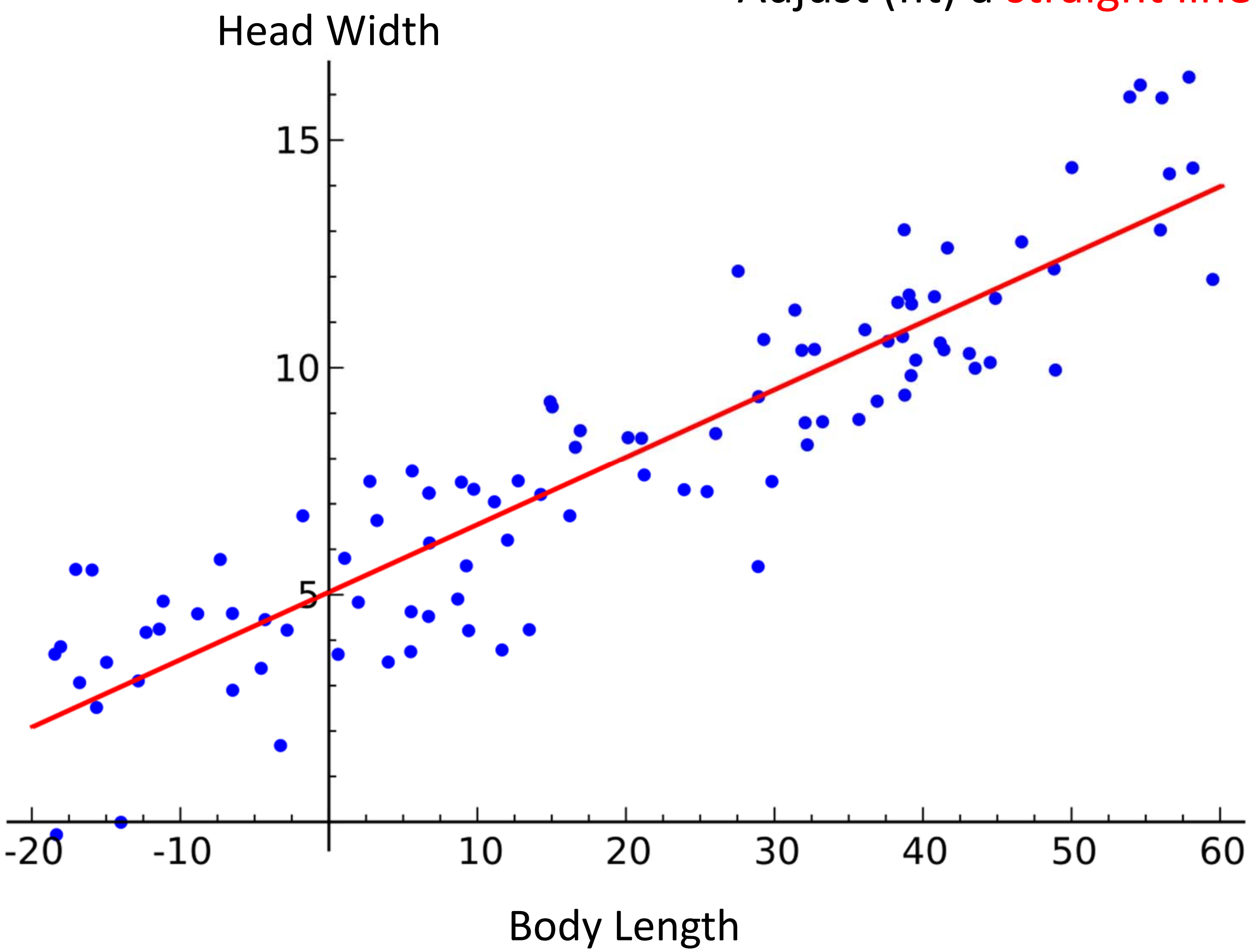
How do linear models make predictions?



LINEAR REGRESSION

1. Data is adjusted to linear regression model:

Adjust (fit) a straight line to our data



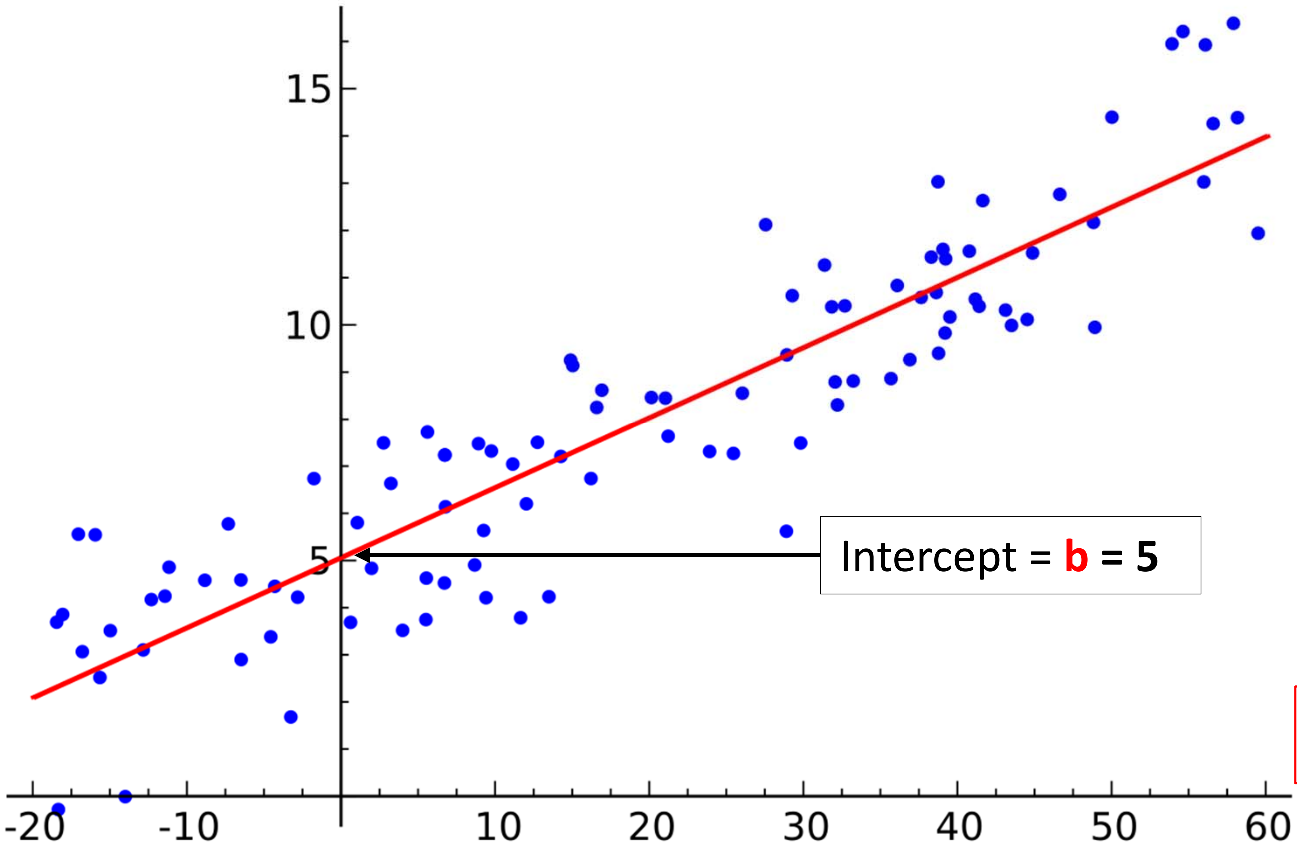
$$y = ax + b$$

LINEAR REGRESSION

2. Calculate expected values:

Predict value of **y** from **x** (we need **a** and **b**)

y = ax + b



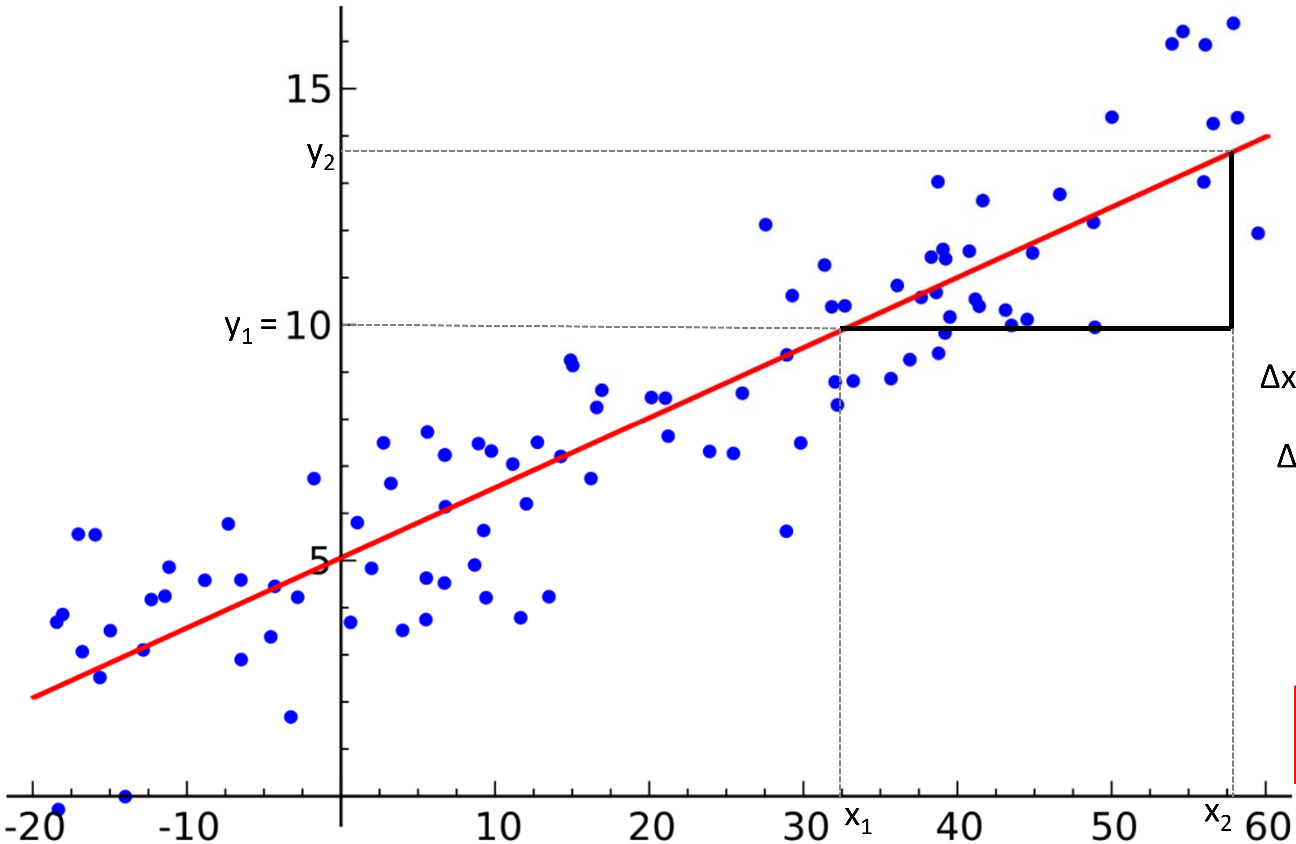
y' = ax + 5

LINEAR REGRESSION

2. Calculate expected values:

Predict value of **y** from **x** (we have **b,** still need **a**)

$y = ax + b$



a = **Slope** = Δy / Δx

$\Delta x = x_2 - x_1 = 57.65 - 32.45 = 25.2$

$\Delta y = y_2 - y_1 = 13.78 - 10 = 3.78$

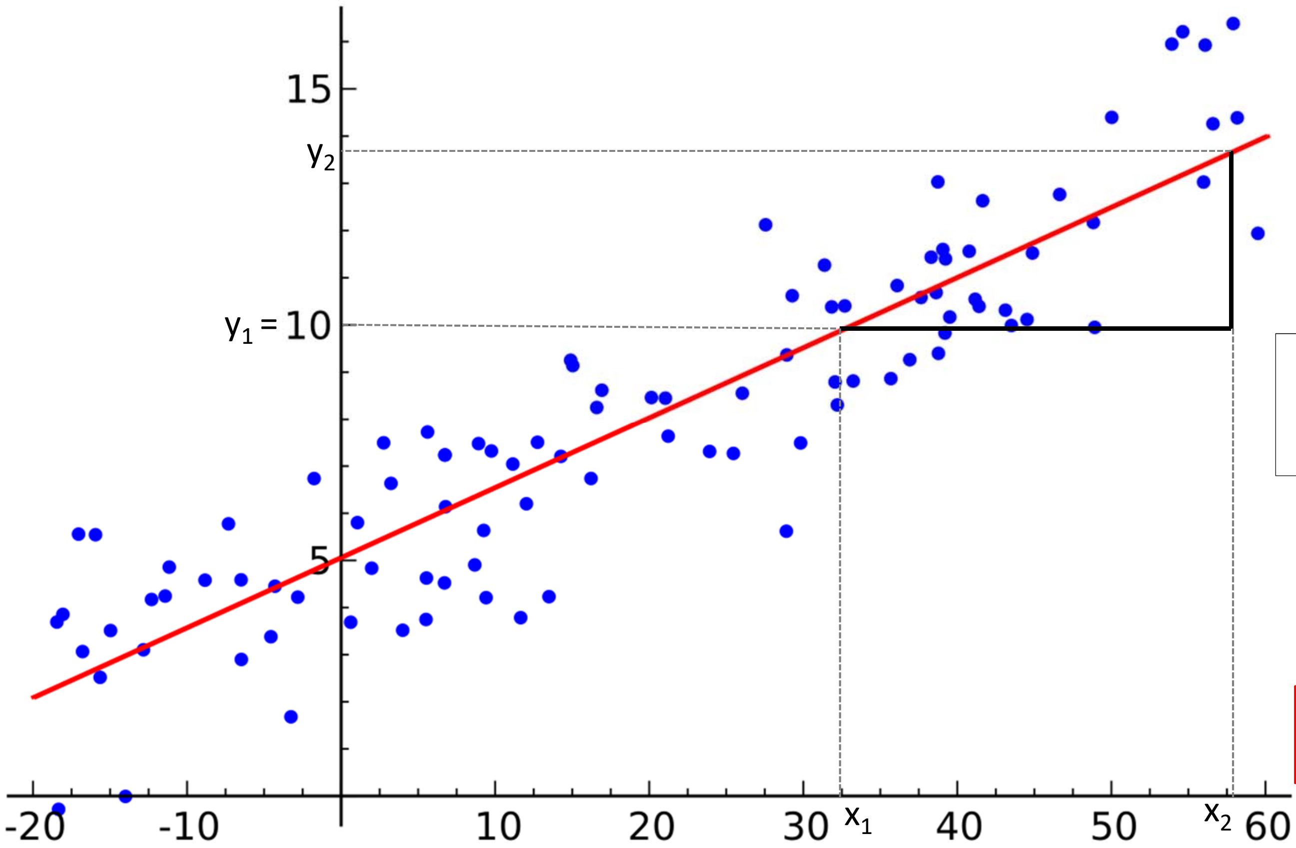
$y' = ax + 5$

LINEAR REGRESSION

2. Calculate expected values:

Predict value of **y** from **x** (we have **b,** still need **a**)

y = ax + b



a = **Slope** = Δy / Δx

Slope = Δy / Δx =
3.78 / 25.2 = **a** = **0.15**

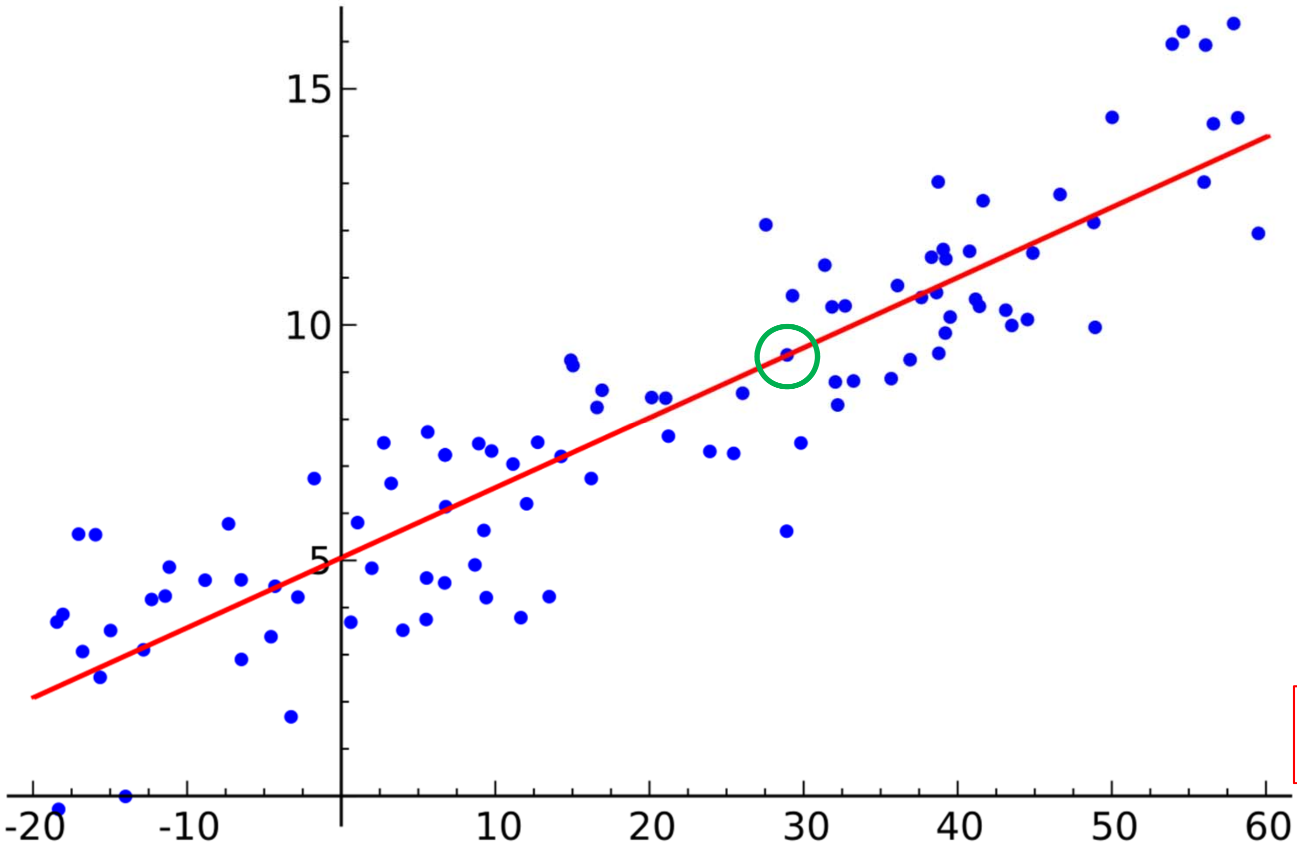
y' = 0.15x + 5

LINEAR REGRESSION

2. Calculate expected values:

Now we can predict value of **y** from **x**

$\mathbf{y = 0.15x + 5}$

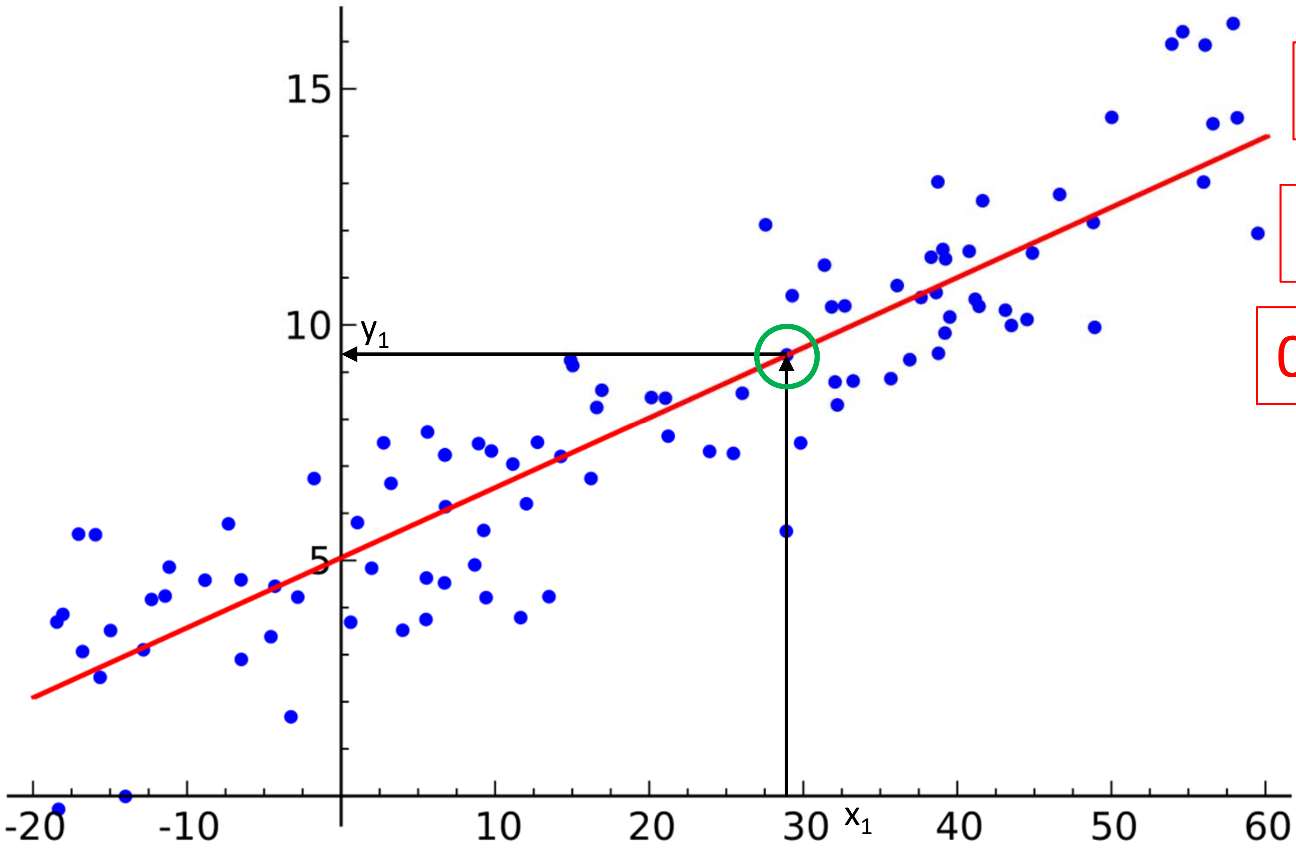


$\mathbf{y' = 0.15x' + 5}$

LINEAR REGRESSION

2. Calculate expected values:

Now we can predict value of **y** from **x**



$0.15\mathbf{x} + 5 = \mathbf{y}$

$0.15 * \mathbf{29} + 5 = \mathbf{y_1}$

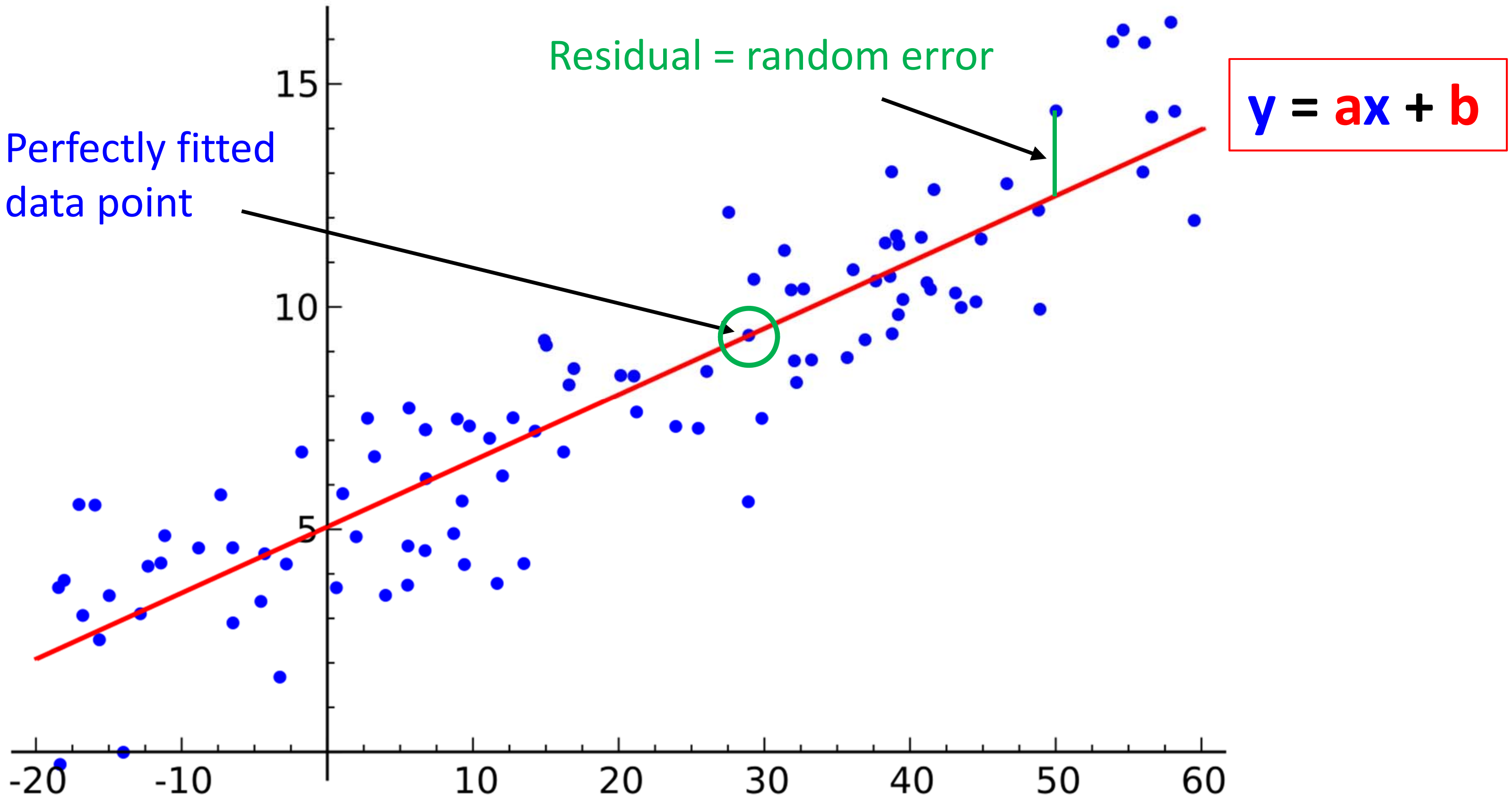
$0.15 * \mathbf{29} + 5 = \mathbf{9.35}$

Specific case for a well adjusted point

LINEAR REGRESSION

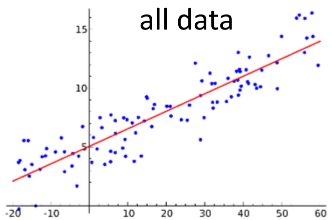
3. Compare expected values with real values:

difference (expected – observed) = error

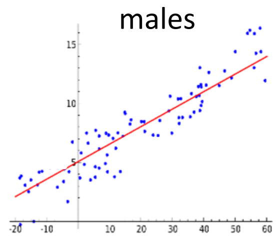


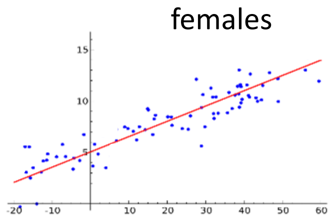
MIND SETUP FOR LINEAR MODELS

Linear models first adjust all data and predict values



Then adjust again the data but **for each group independently** and predict values again.



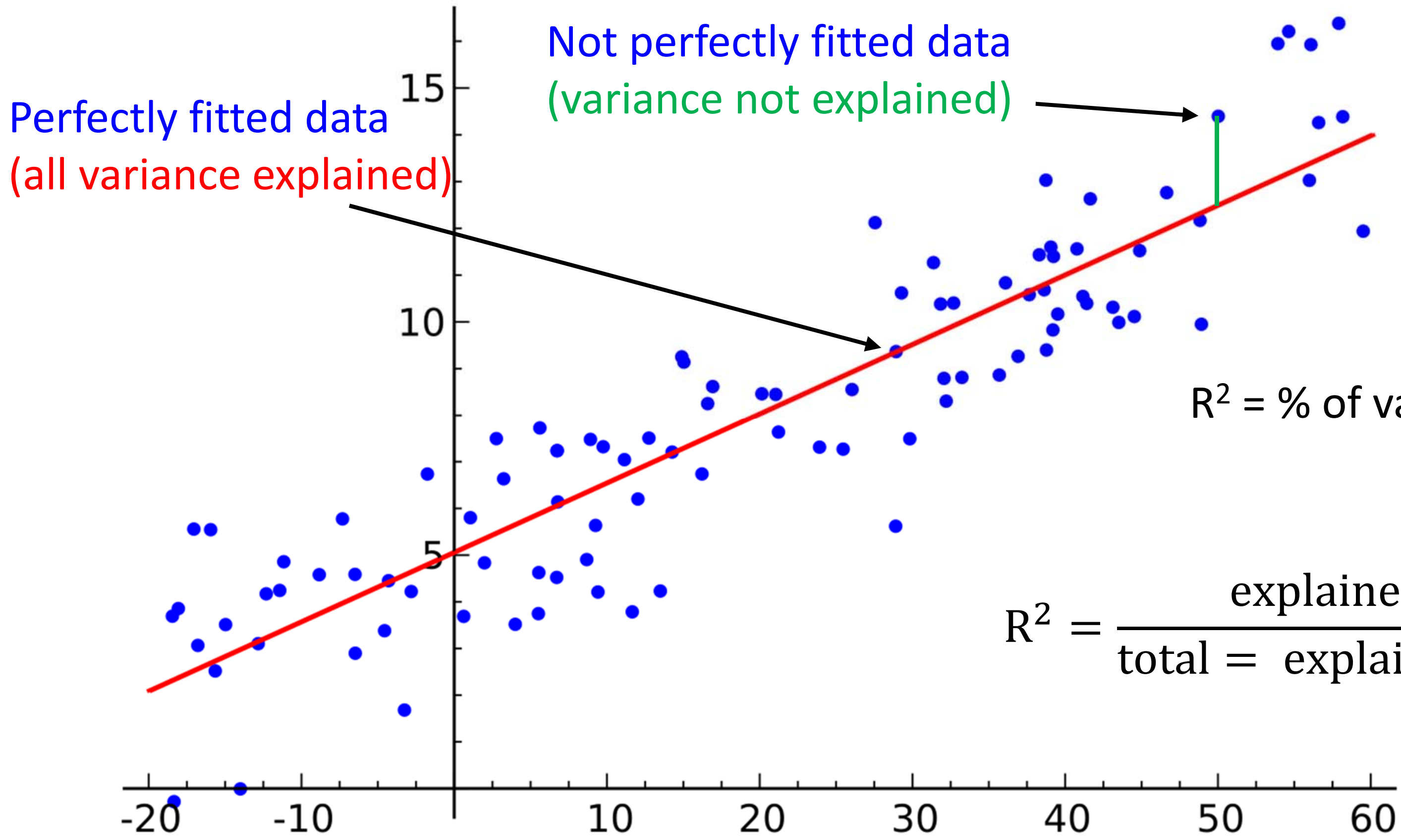


Then compares:
Which prediction is better the one from the model with all data or from the model adjusted separately for each group?

Better = smaller "error"

LINEAR REGRESSION

Linear regression model: Goodness of fit



R^2 = % of variance explained

$$R^2 = \frac{\text{explained by the model}}{\text{total} = \text{ explained} + \text{not explained}}$$

General Linear Models

We already know how to run one **linear model** in ***R***: ANOVA

aov (response~explanatory, data)

aov (Head Height ~ Group, data=dataset)

We could tell ***R*** to run a **general linear model**

Linear model -> **lm()**

- Will assume **normality**

- Will choose the analysis according to our data
  - Response is continuous; Explanatory is categorical/discrete -> **ANOVA**
  - Response is continuous; Explanatory is continuous -> **linear regression**

General Linear Models

lm ()

- Needs to be normal distributed
- Only one continuous response variable
- One or more (few) explanatory variables (categorical or continuous)

Data Distribution	Response variable	Explanatory variable	# Predictors	Test
Normal	*continuous*	*discrete*	*one*	*One-way ANOVA*
Normal	*continuous*	*discrete*	*multiple*	*Multi-way ANOVA*
Normal	*continuous*	*continuous*	*one*	*Linear regression*
Normal	*continuous*	*continuous*	*multiple*	*Multiple regression*
Normal	*continuous*	*discrete & continuous*	*multiple*	*ANCOVA*

General Linear Models

lm (response ~ explanatory, data=dataset)

lm (Head Width~ Group)	->	One-way ANOVA
*lm (Head Width ~ Pop*Sex)*	->	Multi-way ANOVA
lm (Head Width ~ Body Length)	->	Linear Regression
*lm (Head Width ~ Jaw Length*Body Length)*	->	Multiple Regression
*lm (Head Width ~ Body Length *Group)*	->	ANCOVA

Data Distribution	Response variable	Explanatory variable	# Predictors	Test
Normal	*continuous*	*discrete*	*one*	*One-way ANOVA*
Normal	*continuous*	*discrete*	*multiple*	*Multi-way ANOVA*
Normal	*continuous*	*continuous*	*one*	*Linear regression*
Normal	*continuous*	*continuous*	*multiple*	*Multiple regression*
Normal	*continuous*	*discrete & continuous*	*multiple*	*ANCOVA*

GENERAL LINEAR MODELS

Agrupation of Linear Models

- Handle only continuous response variables
- Handle only data with Normal distributions
- Only one type of regression (linear)
- Only one response variable
- Handle a few explanatory variables
- Significance checked by maximum likelihood (p-value)

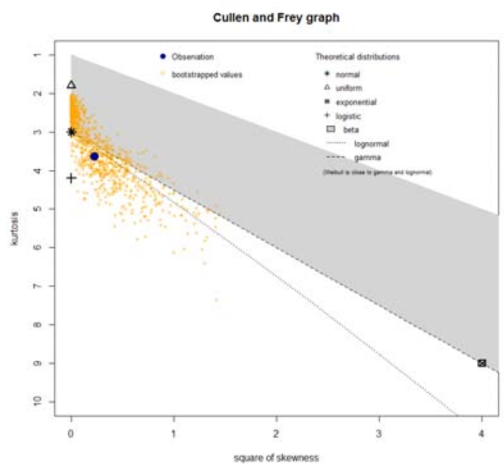
OTHER DISTRIBUTIONS?

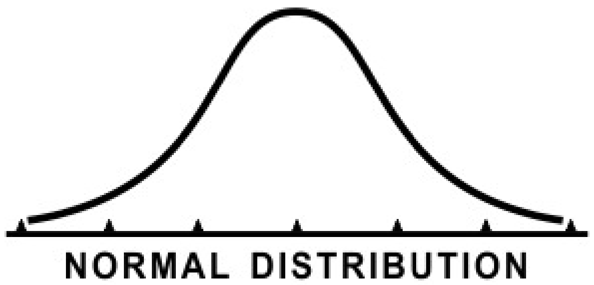
If distribution does not fit the Normal distribution:

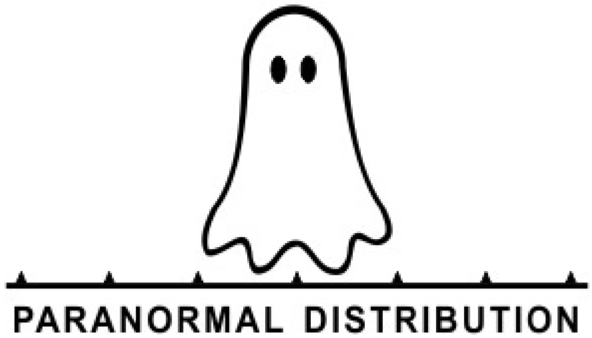
- transformation of dataset (log?)

  OR

- use of low-power non parametric tests (Kruskal-Wallis)

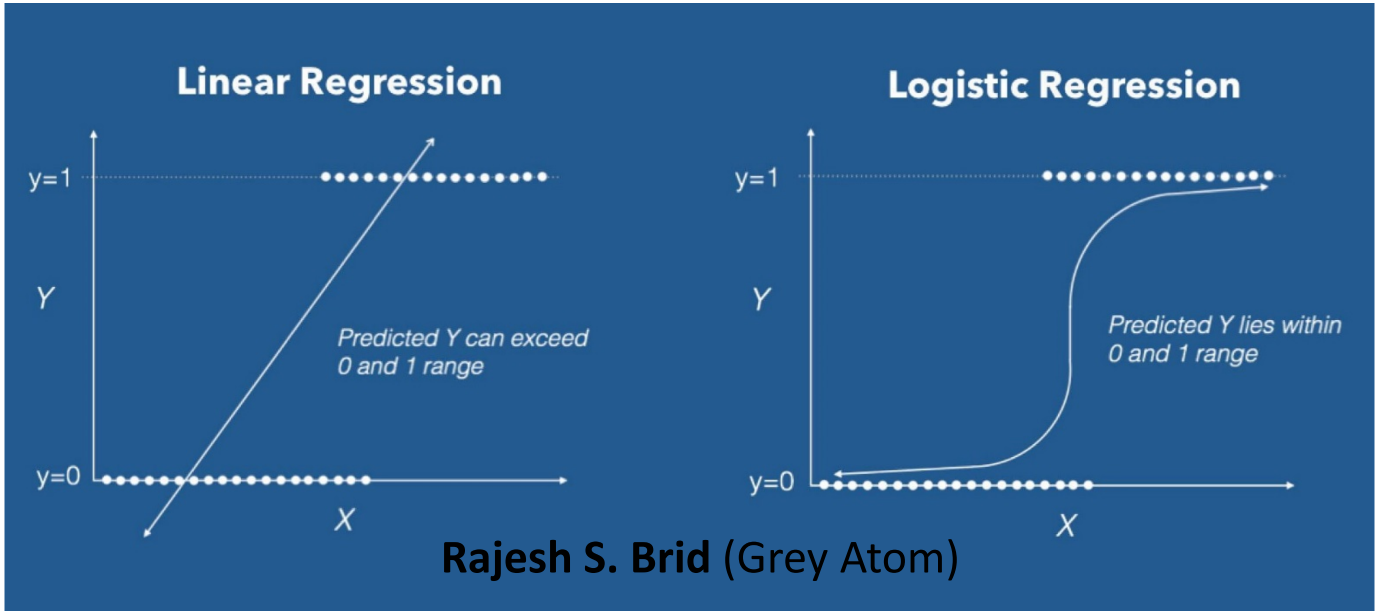






OTHER REGRESSIONS?

- Logistic regression (1/0; yes/no; male/female)
- Poisson regression (discrete and ordinal data)



Rajesh S. Brid (Grey Atom)

GENERAL LINEAR MODELS

Agrupation of Linear Models

- Handle only continuous response variables
- Handle only data with Normal distributions
- Only one type of regression (linear)
- Only one response variable
- Handle a few explanatory variables
- Significance checked by maximum likelihood (p-value)

GENERALI**ZED** LINEAR MODELS

Generalization of **Genera**l Linear Models

- Response variable can be from continuous to categorical
- Handle data with Normal or **other** distributions
- Adjust model to linear and **other** regressions
- Handle many variables at the same time
  and their interactions (N/10)

GENERAL**IZED** LINEAR MODELS

Regressions:

Linear, multivariate, logistic, Poisson's

Distributions:

Normal, Poisson, Binomial, Multinomial

Significance:

Maximum likelihood, Bayesian,

and least squares

Data:

Continuous, Count, Probability,

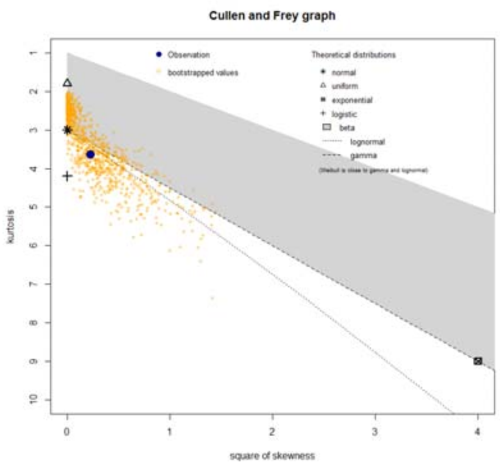
Frequency, Binary, etc.

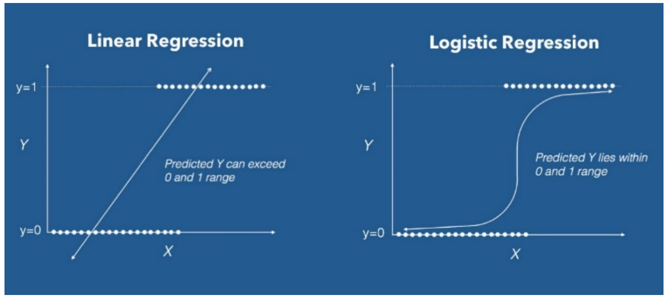
GLM LINK FUNCTIONS

Link function links the explanatory variables

Link function for linear regression $\boxed{y = ax + b}$

GLM handles many types of data, regressions, and models
We need to tell R which kind of link function will need to apply





GLM LINK FUNCTIONS

Don't need to transform data to fit it in a distribution

You tell GLM to which **family** (type of data) belongs your data and it will choose the right link function to process it!

Which Link function?

Tell GLM to which FAMILY belongs our data:

Family is a combination of data characteristics and distributions: Continuous / categorical, includes zeros (0), does not include zeros, has negative values, only positive, type of distribution, etc.

Rules of thumb to choose "family"

DATA	FAMILY
Normal Distribution →	Gaussian
Positive continuous data, no zeros →	Gamma
Gamma distribution	Gamma
Binary data (y/n) →	Binomial
Proportions (3:1) →	Binomial
Logistic distribution →	Binomial
Counts →	Poisson
Categorical (SD= $\bar{y}$) →	Poisson
Poisson distribution →	Poisson
Log Normal distribution →	Poisson
Other categorical →	Negative binomial

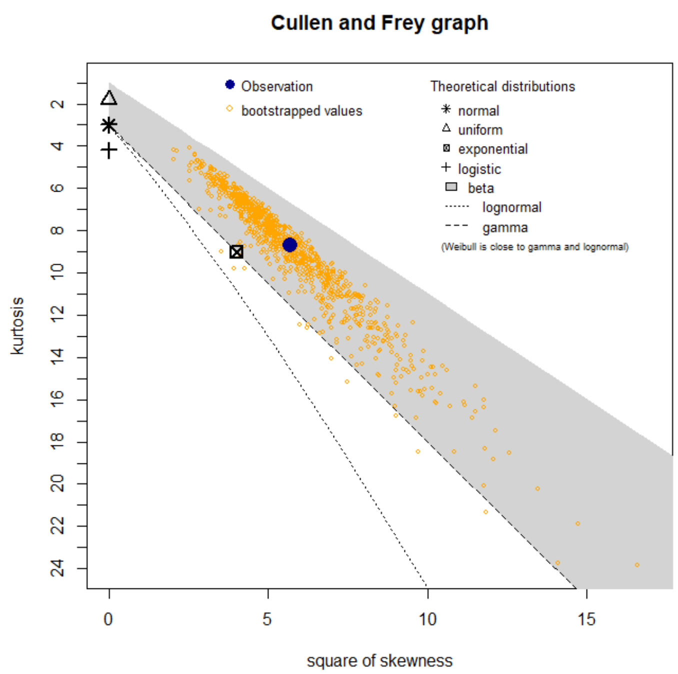
If we doubt which family fits better

fitdistrplus()

Check the fit of some typical distributions with **descdist()** plot

Distribution		Family
Normal	→	Gaussian
Logistic	→	Binomial
Gamma	→	Gamma
LogNormal	→	Poisson

Check other specific distributions with **fitdist()**



Running GLMs

1. Choose variables, choose a family
2. If not sure, or want to be extra sure:

   fitdistr()

   descdist()
3. run the analysis and check if model is well fitted

   glm (response(s)~predictor(s), family, data)

Understanding

OUTPUTS

GENERALIZED LINEAR MODELS (open RStudio)

Understanding GLMs output

```
> summary(glm1)
Call:
glm(formula = Sex ~ Pop, family = "binomial", data = morelizards)
Deviance Residuals:
     Min       1Q    Median       3Q      Max
  -1.317   -1.317    1.044     1.044    1.127

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.228e-01  2.026e-01   1.593    0.111
   PK        7.837e-16  2.865e-01   0.000    1.000
   PM       -2.026e-01  2.849e-01  -0.711    0.477

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 410.39  on 297  degrees of freedom
AIC: 416.39
Number of Fisher Scoring iterations: 4
```

Understanding GLMs output

```
> summary(glm1)
Call:
glm(formula = Sex ~ Pop, family = "binomial", data = morelizards)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
  -1.317    -1.317     1.044     1.044     1.127
```

HOW WELL ADJUSTED IS THE MODEL?
AIC can be used to compare with other models, but is not an absolute

```
Coeffi
(Intercept)   3.228e-01  2.026e-01   1.593     0.111
   PK         7.837e-16  2.865e-01   0.000     1.000
   PM        -2.026e-01  2.849e-01  -0.711     0.477

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 410.39  on 297  degrees of freedom
AIC: 416.39
Number of Fisher Scoring iterations: 4
```

Understanding GLMs output

```
> summary(glm1)
Call:
glm(formula = Sex ~ Pop, family = "binomial", data = morelizards)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
  -1.317   -1.317    1.044    1.044    1.127

Coeffi

(Inter
   PK
   PM

(Dispe

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 410.39  on 297  degrees of freedom
AIC: 416.39
Number of Fisher Scoring iterations: 4
```

HOW WELL ADJUSTED IS THE MODEL?

AIC can be used to compare with other models, but is not an absolute

Compare how well our data fits each distribution with:

Akaike Information Criterion **(AIC)**

The lower the value the better.

Understanding GLMs output

```
> summary(glm1)
Call:
glm(formula = Sex ~ Pop, family = "binomial", data = morelizards)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
  -1.317    -1.317     1.044     1.044     1.127

Coeffi
(Intercept)  3.228e-01  2.026e-01   1.593    0.111
   PK        7.837e-16  2.865e-01   0.000    1.000
   PM       -2.026e-01  2.849e-01  -0.711    0.477

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 410.39  on 297  degrees of freedom
AIC: 416.39
Number of Fisher Scoring iterations: 4
```

HOW WELL ADJUSTED IS THE MODEL?

There is no R^2 we need to calculate ourselves

Understanding GLMs output

```
> summary(glm1)
Call:
glm(formula = Sex ~ Pop, family = "binomial", data = morelizards)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
  -1.317   -1.317    1.044    1.044    1.127

Coeffi

(Inter
   PK
   PM
```

HOW WELL ADJUSTED IS THE MODEL?

There is no R^2 we need to calculate ourselves

$$R^2 = \frac{\text{Null deviance - Residual deviance}}{\text{Null deviance}} = \frac{411.06 - 410.39}{411.06} = 0.0016$$

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 410.39  on 297  degrees of freedom
AIC: 416.39
Number of Fisher Scoring iterations: 4
```

Understanding GLMs output

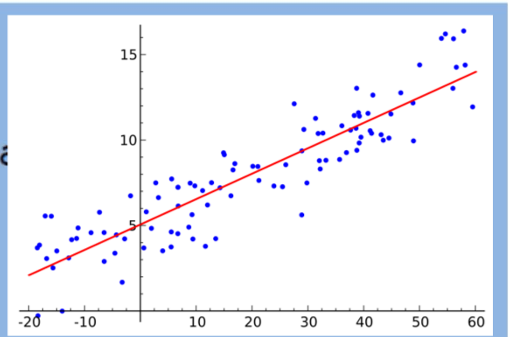
```
> summary(glm1)
Call:
glm(formula = Sex ~ Pop, family = "binomial", data
Deviance Residuals:
     Min        1Q    Median        3Q       Max
  -1.317    -1.317     1.044     1.044     1.127

Coeffi

(Inter
   PK
   PM

(Dispe

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 410.39  on 297  degrees of freedom
AIC: 416.39
Number of Fisher Scoring iterations: 4
```



HOW WELL ADJUSTED IS THE MODEL?

Rule of Thumb:

If: degrees of freedom*2 **<** Residual deviance ⟶ Data are **overdispersed**

297*2 = 594 ≳ 410.39 ⟶ Fit is not great, but data are **not overdispersed ✔**

Understanding GLMs output

```
> summary(glm1)
Call:
glm(formula = Sex ~ Pop, family = "binomial", data = morelizards)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
  -1.317  -1.317    1.044    1.044    1.127

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.228e-01  2.026e-01   1.593    0.111
   PK        7.837e-16  2.865e-01   0.000    1.000
   PM       -2.026e-01  2.849e-01  -0.711    0.477

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 410.39  on 297  degrees of freedom
AIC: 416.39
Number of Fisher Scoring iterations: 4
```

Understanding GLMs output

```
> summary(glm1)
Call:
glm(formula = Sex ~ Pop, family = "binomial", data = morelizards)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
  -1.317    -1.317     1.044     1.044     1.127

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.228e-01  2.026e-01   1.593    0.111
   PK        7.837e-16  2.865e-01   0.000    1.000
   PM       -2.026e-01  2.849e-01  -0.711    0.477

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.06  on 299  degrees of freedom
Residual deviance: 410.39  on 297  degrees of freedom
AIC: 416.39
Number of Fisher Scoring iterations: 4
```

Each "free" group/covariable comparison against the intercept. Interesting, but not super useful

TASKS

https://tinyurl.com/evolecopract --> **morelizards.csv** and **fifth_linear_models.R**

GLM

Libraries: fitdistrplus, car, boot

1. Choose variables (discrete or continuous, all allowed)
2. Plot the distribution: *descdist(variable)*
3. Fit your data to distributions <u>that make sense for your type of data</u> and compare how well they fit

   fit1<-fitdistr(na.omit(variable), "negative binomial")

   AIC (fit1, fit2, ...) #The lowest the better
4. Choose the families that better adjust to your data:
   gaussian, poisson, poisson, Gamma, binomial, negative.binomial
5. **Perform glm()**
   glm(response~explanatory, family=negative.binomial, data=dataset)
6. Check AIC, calculate R^2 and "degrees of freedom vs variance"
7. Perform an ANOVA with the output: *aov(glm_output)*
8. Are there significant effects for the chosen variables?
9. Plot