Multiple regression

[brouwern@gmail.com](mailto:brouwern@gmail.com)

December 5, 2016

# Multiple linear regression

* Also known as "multiple regression"

## Lion data: predicting age from nose coloration

### Preliminaries

Load the data

lions <- read.csv("lion\_age\_by\_pop.csv" )

There are 2 populations, The main Serengeti population and the Ngorogoro crater (sub?) population

summary(lions)

## portion.black age.years population  
## Min. :0.1000 Min. : 1.100 Ngorogoro:10   
## 1st Qu.:0.1475 1st Qu.: 2.175 Serengeti:22   
## Median :0.2650 Median : 3.500   
## Mean :0.3197 Mean : 4.309   
## 3rd Qu.:0.4325 3rd Qu.: 5.850   
## Max. :0.7900 Max. :13.100

In the book they consider *all* of the data combined and ignore the two seperate populations (which are sperate geographically and genetically)

Does the relationship between age and nose-blackness work the same for both populations?

### Plot lion data w/ both populations combined

We will "pool" the data from both populations (that is, ignore population differences)

#### Run regression

#build regression model  
m.pooled <- lm(age.years ~ portion.black, data = lions)

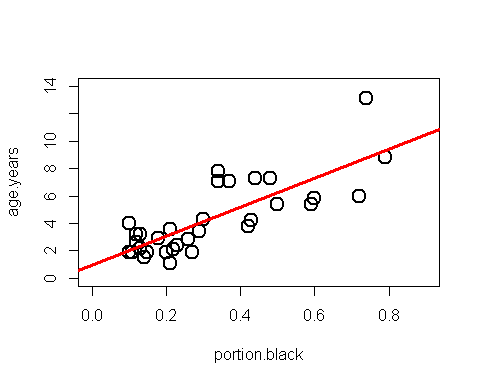
#### Look at summary of regression

summary(m.pooled)

##   
## Call:  
## lm(formula = age.years ~ portion.black, data = lions)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5457 -1.1457 -0.3384 0.9245 4.3426   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.9262 0.5591 1.657 0.108   
## portion.black 10.5827 1.4884 7.110 6.59e-08 \*\*

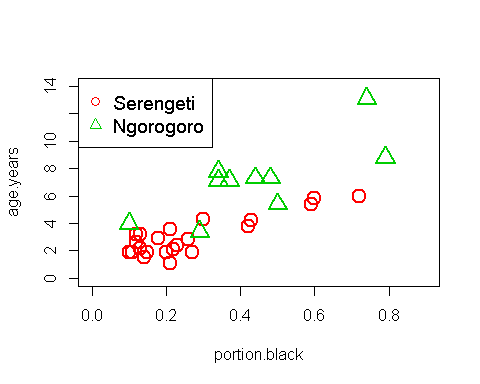
## Multiple R-squared: 0.6276, Adjusted R-squared: 0.6  
## F-statistic: 50.55 on 1 and 30 DF, p-value: 6.59e-08

#### Plot pooled data



#To Do: add equations for line

### Plot lion data separated by population



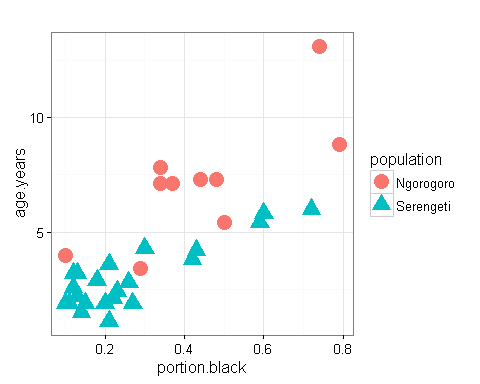
#### What model best describes the data: 1 line or 2?

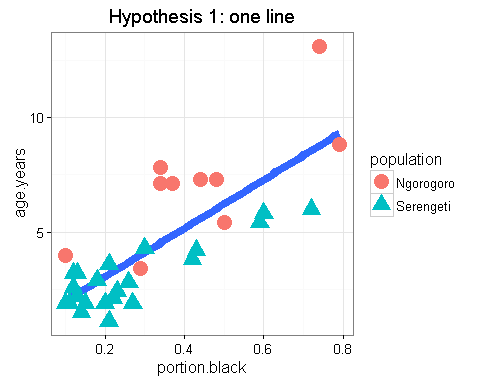
##### Hypothesis 0: there is no relationship

THe null hypothesis Ho is that there is no relationship between how black the nose is and age. We can seee that there is a relationshop and will skip this Ho.

##### Hypothesis 1: the populations are the same

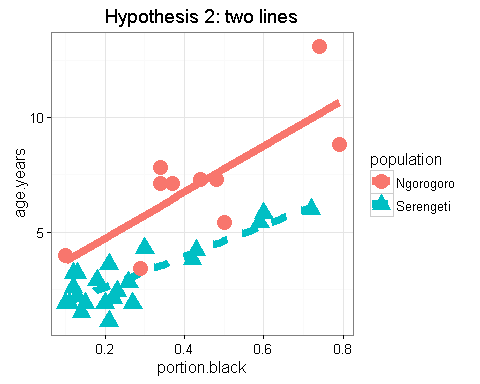
The data can be modeled with a single regression line. Any differene between the popuatliosn due to random noise.





##### Hypothesis 2: the populations are different

*The populations have different biological/ecological processes going on that make the relationship between age and nose blackness depend on the population* We say there is an **INTERACTION** between population and portion.black *stated another way: The relationship between nose blackness and age is different between the two populations* Therefore, the slope of the lines are different \*(THis include the possiblity that there is no relationship in one population (slope = 0) and is a relationship in the other (slope >0) )



#### Run multiple regression model

*We indicate in* ***interaction*** *in R using the multiple symbol "*"