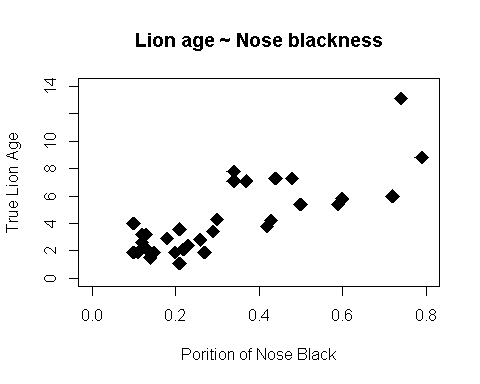
Regression Example

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

November 30, 2016

### Lion data Table 17.1-1, pg 542

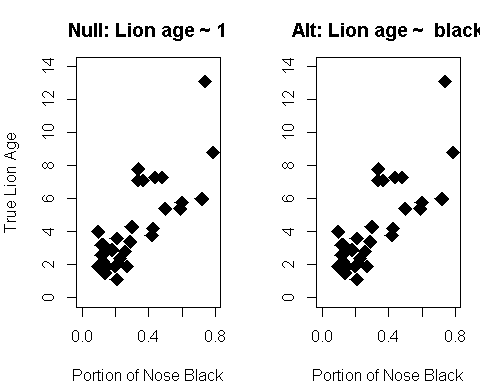


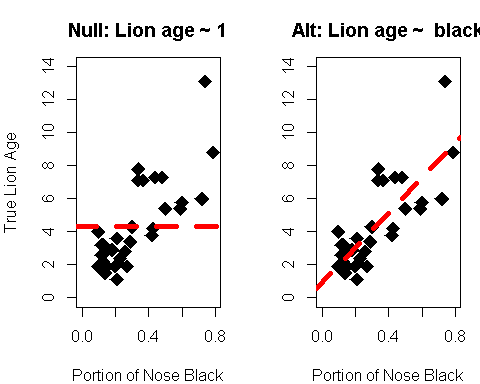
### Regression of Lion Data

Fit Regression models

## Null model: no relationship between how black nose is and true age  
lion.null <- lm(age.years ~ 1, # What does "~ 1 " tell R?  
 data = lions)  
  
  
## ALt model: Age can be predicted by portion of nose that is black  
lion.alt <- lm(age.years ~ portion.black,   
 data = lions)

### Visualize the hypotheses posed by the models

Can you draw the hypotheses? 



### Regression Model Output

#### Standard way of getting output with summary() command

##   
## 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 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.66 on 30 degrees of freedom  
## Multiple R-squared: 0.6276, Adjusted R-squared: 0.6152   
## F-statistic: 50.55 on 1 and 30 DF, p-value: 6.59e-08

This gives a lot of somewhat cluttered output

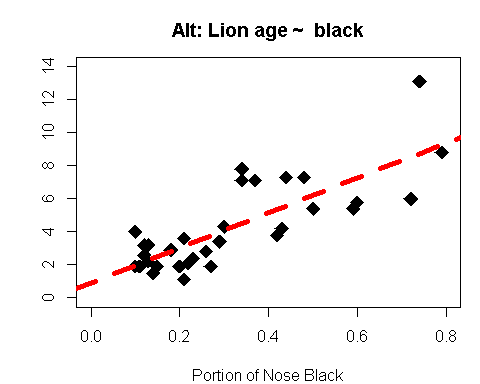
#### Just look at regression coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.926 0.559 1.657 0.108  
## portion.black 10.583 1.488 7.110 0.000

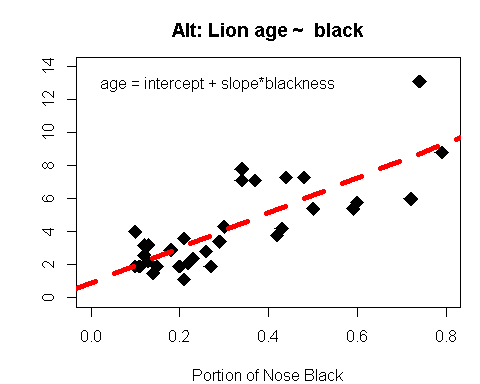
#### Questions you should be able to answer:

* Why is one of the p-values equal to zero?
* Which one is the slope of the line?
* What is the biological meaning of a non-significant intercept?

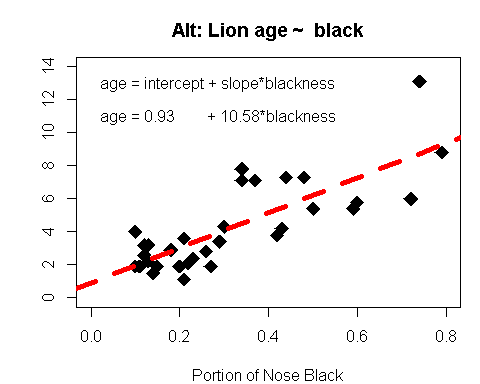
### The regression model tells us the equation of the line



## (Intercept) portion.black   
## 0.9262292 10.5826652



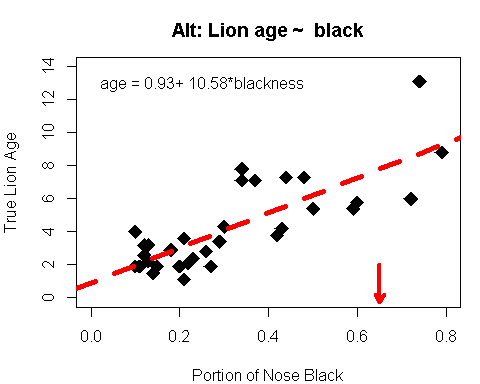
## (Intercept) portion.black   
## 0.9262292 10.5826652



### Predictions from a model

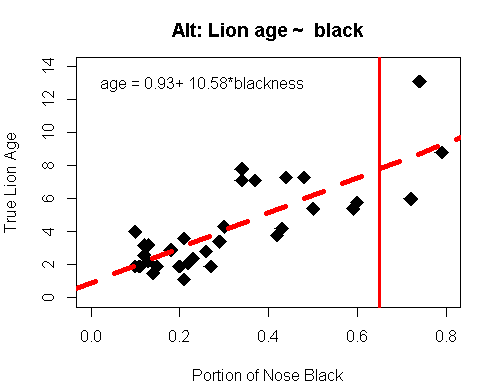
* Say we take a picture of a lion and determine that the portion of its nose that is blck is 0.65 (65%)
* What is the *predicted* age of the lion

## (Intercept) portion.black   
## 0.9262292 10.5826652

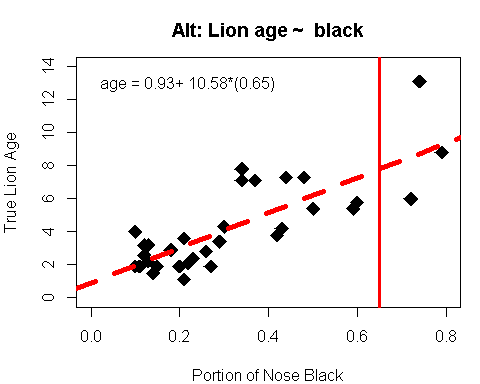


Points along the regression lines can be thought of as predictions for each value of x if we found a new lion with the value of x.

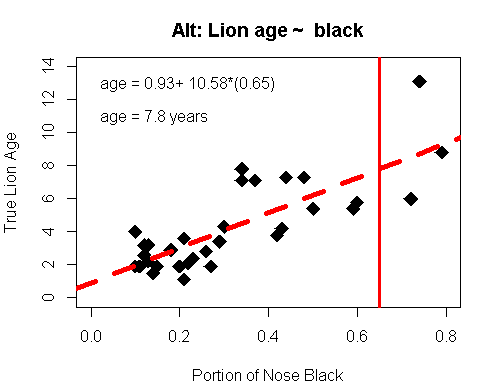
## (Intercept) portion.black   
## 0.9262292 10.5826652



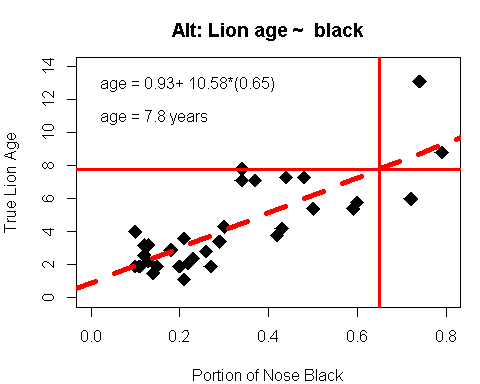
## (Intercept) portion.black   
## 0.9262292 10.5826652



## (Intercept) portion.black   
## 0.9262292 10.5826652



## (Intercept) portion.black   
## 0.9262292 10.5826652



### Getting predictions in R

* Uses predict() function
* I have written a function called easy.predict() to help with this

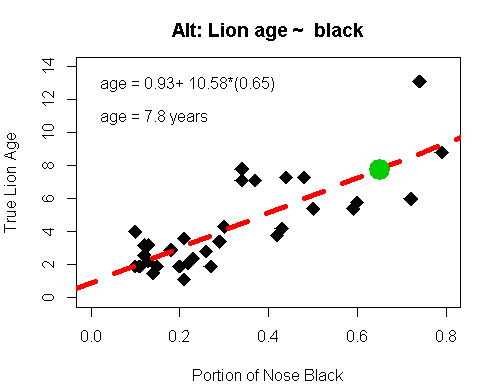
easy.predict <- function(model,datum){  
 x.var.name <- names(model$coefficients)[[2]]  
new.df <- data.frame(datum)  
names(new.df) <- x.var.name  
  
 predict(model, newdata = new.df)[[1]]  
}

### Get prediction for portion blakc = 0.65

easy.predict(lion.alt, datum = 0.65)

## [1] 7.804962

### Plot the prediciton



### Is regression appropriate for these data

Do we violate the **assumptions** of the regression model?

#### Regression assumptions

Shorthand versions (You MUST know these from memory) \* 1)Random sampling \* 2)Variability around regression line is approximately normal \* 3)Variance in y does not change as x-changes

## Checking Assumption: Model diagnostics

* This can be thought of as "Analysis of the residuals"
* residuals = (real data) - (predictions from model)

You get the residuals in R with the resid() command

#Get the residuals  
model.residuals <- resid(lion.alt)  
  
#Look at first 12  
model.residuals[1:12]

## 1 2 3 4 5 6   
## -2.04858891 -0.90780235 -0.19032239 -0.10197570 0.40385095 0.89802430   
## 7 8 9 10 11 12   
## 1.00385095 0.06889104 -0.96024222 -1.15441556 -1.14276226 -0.08449574

Note that they can be both positive AND negative. Why is that?

### Assumption 1: random sampling

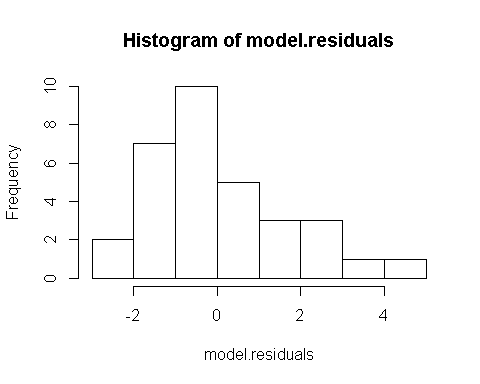
Frequently violated, especially with regression...

### Assumption 2: normality

#### The most basic model diagnostic: are residuals normal-ish distributed

use the hist() command

hist(model.residuals)



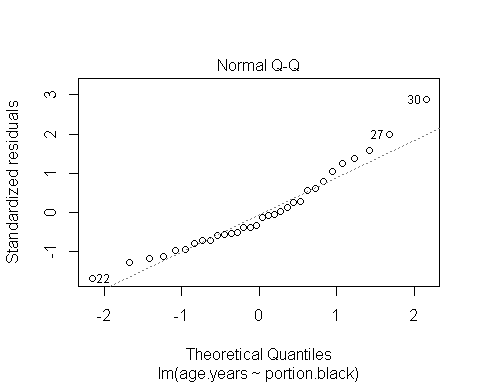
* Kinda look skewed...
* ...but not horrible
* Normality is NOT a particularly important assumption

### Assumption 2: Normality

#### Better look at normality: qqplot in R

"qq" = "quantile quantile"

plot(lion.alt, which = 2)



### 3: Constant variance

plot(lion.alt, which = 1)

