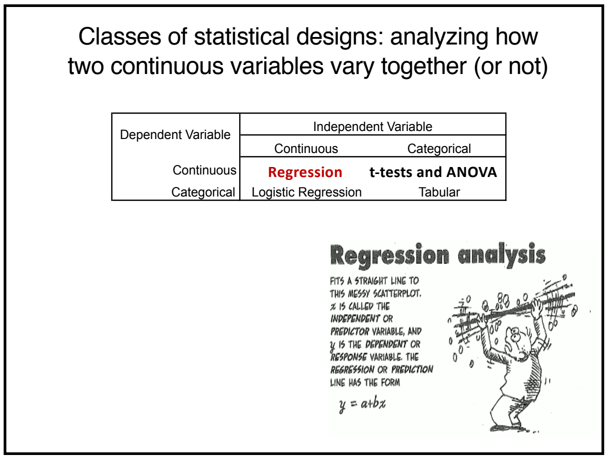
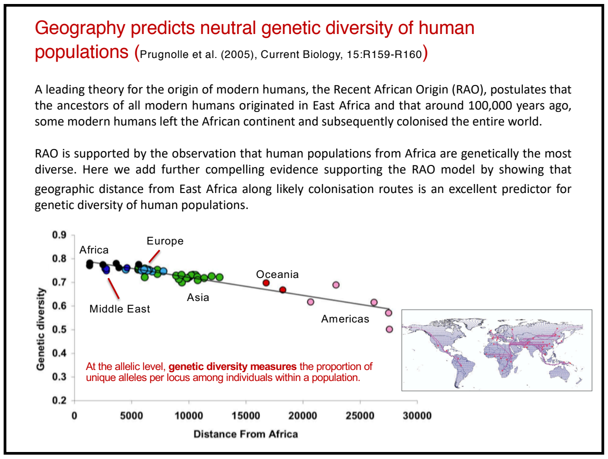


Classes of statistical designs: analyzing how two continuous variables vary together (or not)

Dependent Variable	Independent Variable	
	Continuous	Categorical
Continuous	**Regression**	**t-tests and ANOVA**
Categorical	Logistic Regression	Tabular

Regression analysis
FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE INDEPENDENT OR PREDICTOR VARIABLE, AND y IS THE DEPENDENT OR RESPONSE VARIABLE. THE REGRESSION OR PREDICTION LINE HAS THE FORM
y = a+bx

1



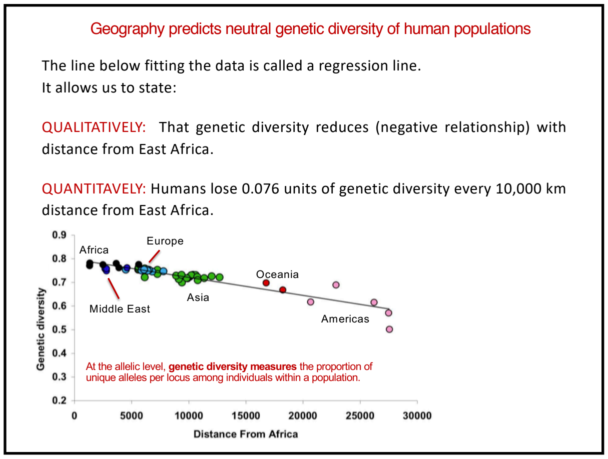
Geography predicts neutral genetic diversity of human populations (Prugnolle et al. (2005), Current Biology, 15:R159-R160)

A leading theory for the origin of modern humans, the Recent African Origin (RAO), postulates that the ancestors of all modern humans originated in East Africa and that around 100,000 years ago, some modern humans left the African continent and subsequently colonised the entire world.

RAO is supported by the observation that human populations from Africa are genetically the most diverse. Here we add further compelling evidence supporting the RAO model by showing that geographic distance from East Africa along likely colonisation routes is an excellent predictor for genetic diversity of human populations.

At the allelic level, **genetic diversity measures** the proportion of unique alleles per locus among individuals within a population.

2



Geography predicts neutral genetic diversity of human populations

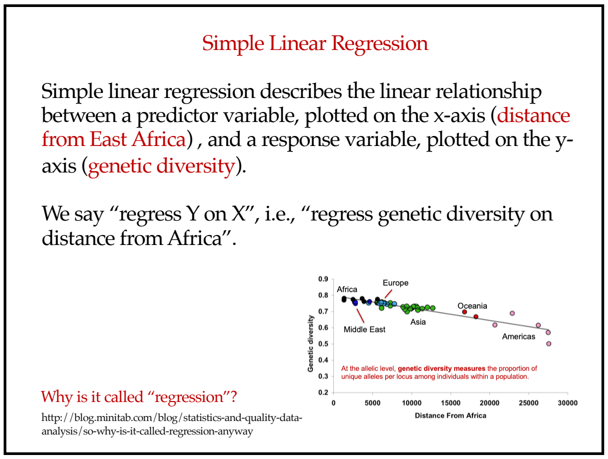
The line below fitting the data is called a regression line.
It allows us to state:

QUALITATIVELY: That genetic diversity reduces (negative relationship) with distance from East Africa.

QUANTITAVELY: Humans lose 0.076 units of genetic diversity every 10,000 km distance from East Africa.

At the allelic level, **genetic diversity measures** the proportion of unique alleles per locus among individuals within a population.

3



Simple Linear Regression

Simple linear regression describes the linear relationship between a predictor variable, plotted on the x-axis (distance from East Africa) , and a response variable, plotted on the y-axis (genetic diversity).

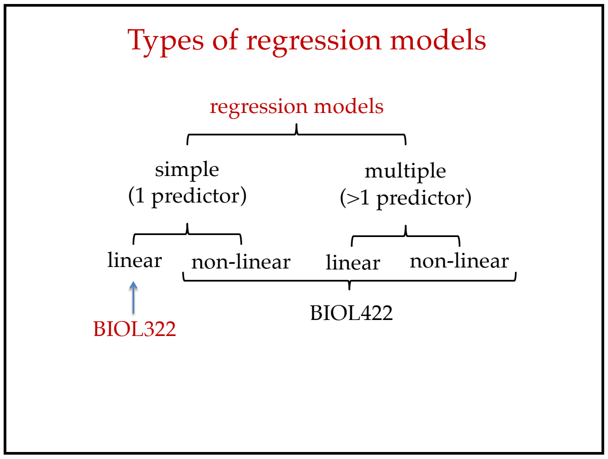
We say "regress Y on X", i.e., "regress genetic diversity on distance from Africa".

At the allelic level, **genetic diversity measures** the proportion of unique alleles per locus among individuals within a population.

Why is it called "regression"?

http://blog.minitab.com/blog/statistics-and-quality-data-analysis/so-why-is-it-called-regression-anyway

4



Types of regression models

regression models

simple (1 predictor)

multiple (>1 predictor)

linear

non-linear

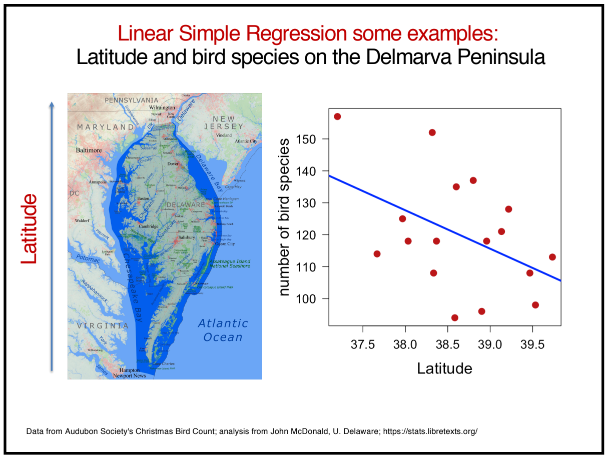
linear

non-linear

BIOL322

BIOL422

5



Linear Simple Regression some examples:
Latitude and bird species on the Delmarva Peninsula

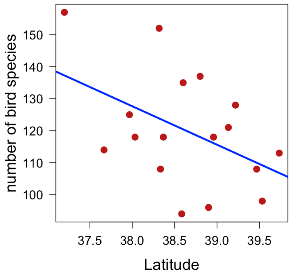
Data from Audubon Society's Christmas Bird Count; analysis from John McDonald, U. Delaware; https://stats.libretexts.org/

6

Linear Simple Regression some examples:
Latitude and bird species on the Delmarva Peninsula

QUALITATIVELY: The number of bird species decreases with Latitude.

QUANTITAVELY: Sites lose 12.04 species every 1º Latitude.



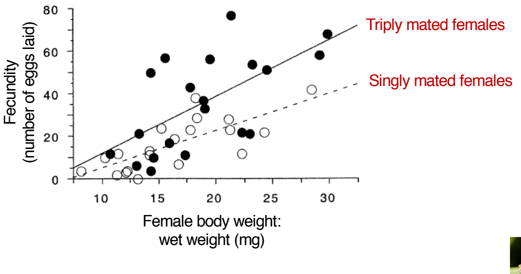
Data from Audubon Society's Christmas Bird Count; analysis from John McDonald, U. Delaware; https://stats.libretexts.org/

7

Fitness advantage from nuptial gifts in female fireflies

Rooney & Lewis (2002), Ecological Entomology, 27:373-377.

During mating, males of both species transfer a single spermatophore that undergoes digestion over several days in a specialized structure within the female reproductive tract.



8

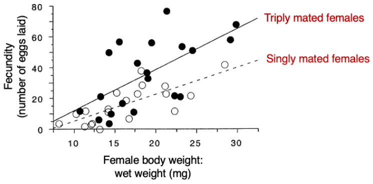
Fitness advantage from nuptial gifts in female fireflies

Rooney & Lewis (2002), Ecological Entomology, 27:373-377.

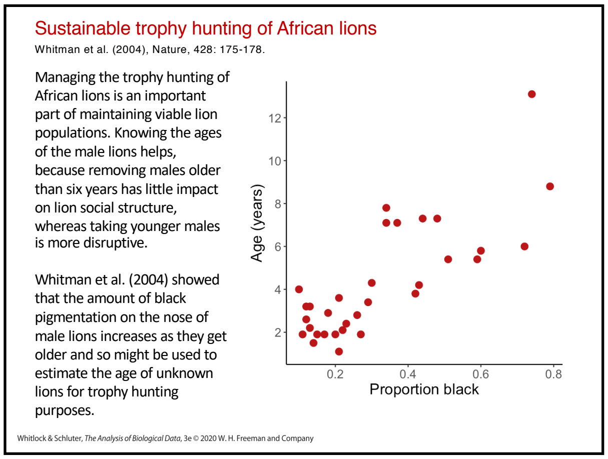
QUALITATIVELY: Fecundity increases with female body weight.

QUANTITAVELY: Triply mated females produce 2.7 eggs per mg of body weight whereas singly mated females produce (less) 1.8 eggs per mg of body weight.

RESEARCH CONCLUSION: Females that receive more nuptial gifts (triply *versus* singly mated) increase their egg production.



9



Sustainable trophy hunting of African lions

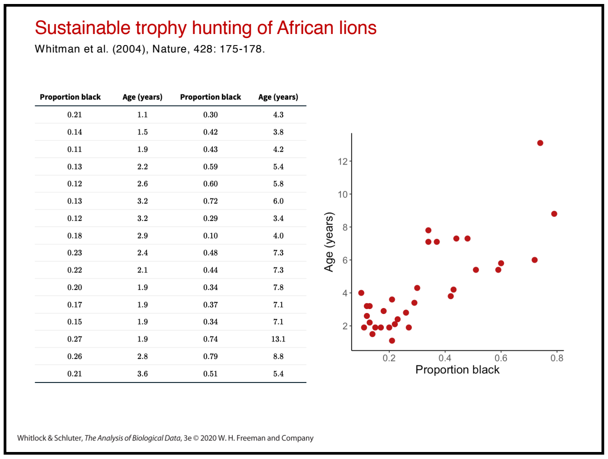
Whitman et al. (2004), Nature, 428: 175-178.

Managing the trophy hunting of African lions is an important part of maintaining viable lion populations. Knowing the ages of the male lions helps, because removing males older than six years has little impact on lion social structure, whereas taking younger males is more disruptive.

Whitman et al. (2004) showed that the amount of black pigmentation on the nose of male lions increases as they get older and so might be used to estimate the age of unknown lions for trophy hunting purposes.

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

10



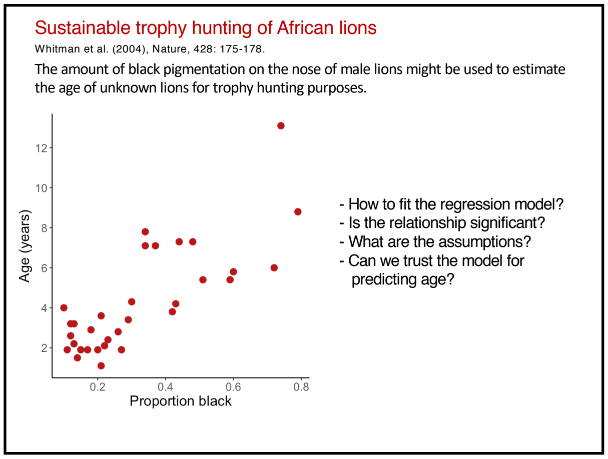
Sustainable trophy hunting of African lions

Whitman et al. (2004), Nature, 428: 175-178.

Proportion black	Age (years)	Proportion black	Age (years)
0.21	1.1	0.30	4.3
0.14	1.5	0.42	3.8
0.11	1.9	0.43	4.2
0.13	2.2	0.59	5.4
0.12	2.6	0.60	5.8
0.13	3.2	0.72	6.0
0.12	3.2	0.29	3.4
0.18	2.9	0.10	4.0
0.23	2.4	0.48	7.3
0.22	2.1	0.44	7.3
0.20	1.9	0.34	7.8
0.17	1.9	0.37	7.1
0.15	1.9	0.34	7.1
0.27	1.9	0.74	13.1
0.26	2.8	0.79	8.8
0.21	3.6	0.51	5.4

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

11



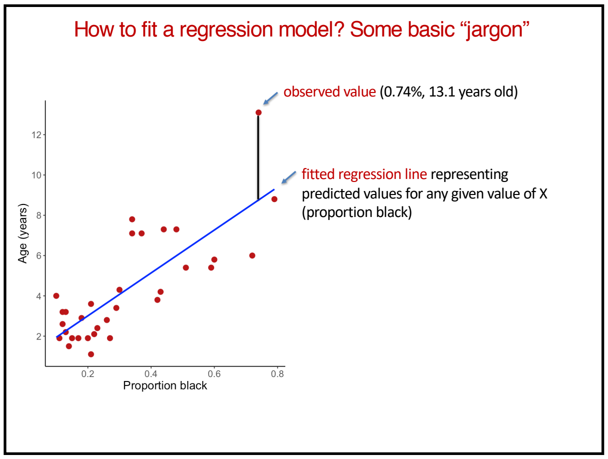
Sustainable trophy hunting of African lions

Whitman et al. (2004), Nature, 428: 175-178.

The amount of black pigmentation on the nose of male lions might be used to estimate the age of unknown lions for trophy hunting purposes.

- How to fit the regression model?
- Is the relationship significant?
- What are the assumptions?
- Can we trust the model for predicting age?

12

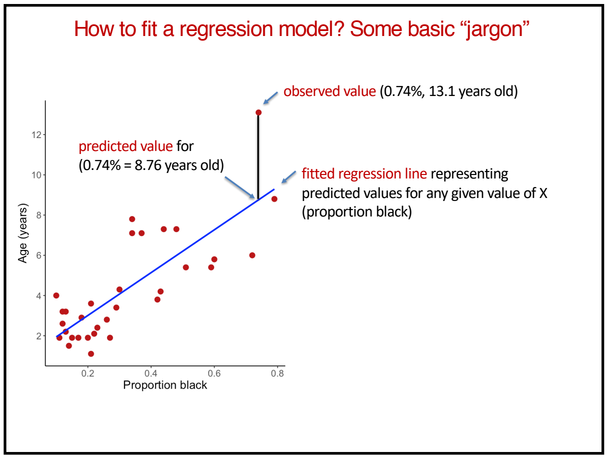


How to fit a regression model? Some basic “jargon”

observed value (0.74%, 13.1 years old)

fitted regression line representing predicted values for any given value of X (proportion black)

13



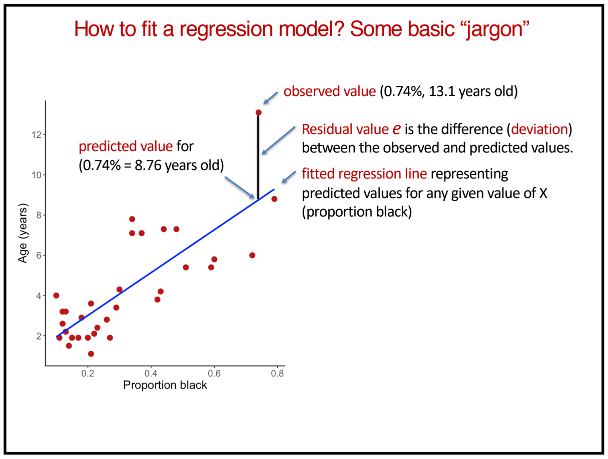
How to fit a regression model? Some basic “jargon”

observed value (0.74%, 13.1 years old)

predicted value for (0.74% = 8.76 years old)

fitted regression line representing predicted values for any given value of X (proportion black)

14



How to fit a regression model? Some basic “jargon”

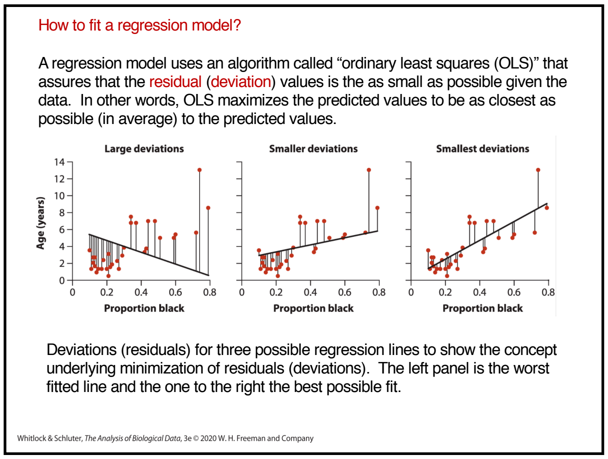
observed value (0.74%, 13.1 years old)

Residual value *e* is the difference (deviation) between the observed and predicted values.

predicted value for (0.74% = 8.76 years old)

fitted regression line representing predicted values for any given value of X (proportion black)

15



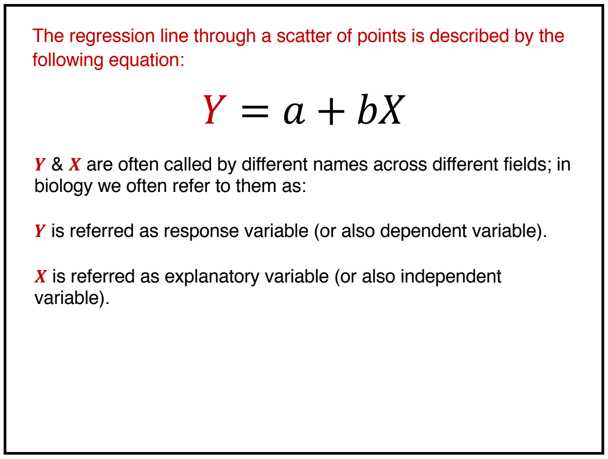
How to fit a regression model?

A regression model uses an algorithm called "ordinary least squares (OLS)" that assures that the residual (deviation) values is the as small as possible given the data. In other words, OLS maximizes the predicted values to be as closest as possible (in average) to the predicted values.

Deviations (residuals) for three possible regression lines to show the concept underlying minimization of residuals (deviations). The left panel is the worst fitted line and the one to the right the best possible fit.

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

16



The regression line through a scatter of points is described by the following equation:

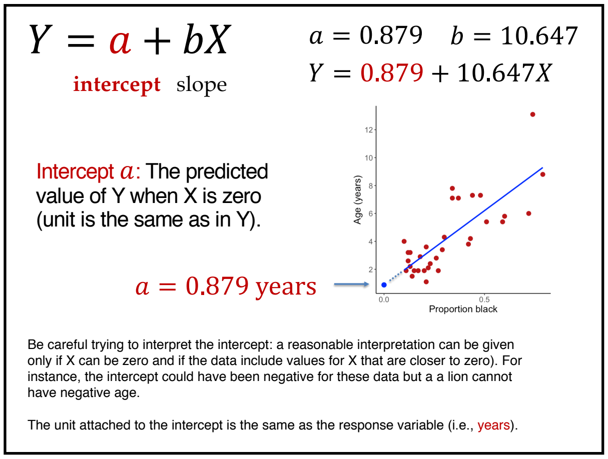
$$Y = a + bX$$

Y & ***X*** are often called by different names across different fields; in biology we often refer to them as:

Y is referred as response variable (or also dependent variable).

X is referred as explanatory variable (or also independent variable).

17



$$Y = a + bX$$

intercept slope

$a = 0.879 \quad b = 10.647$

$Y = 0.879 + 10.647X$

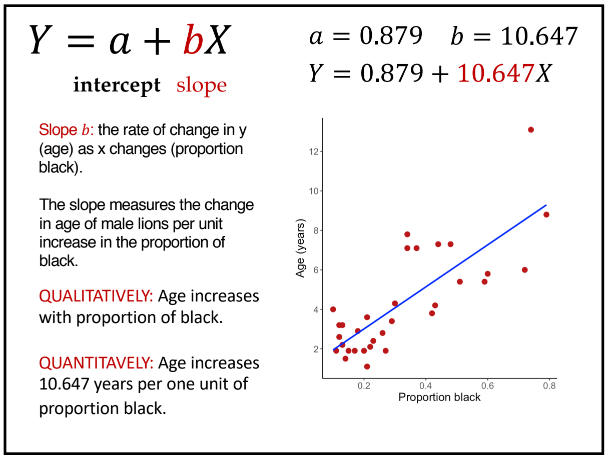
Intercept a: The predicted value of Y when X is zero (unit is the same as in Y).

$a = 0.879$ years

Be careful trying to interpret the intercept: a reasonable interpretation can be given only if X can be zero and if the data include values for X that are closer to zero). For instance, the intercept could have been negative for these data but a a lion cannot have negative age.

The unit attached to the intercept is the same as the response variable (i.e., years).

18



$Y = a + bX$

intercept slope

$a = 0.879 \quad b = 10.647$

$Y = 0.879 + 10.647X$

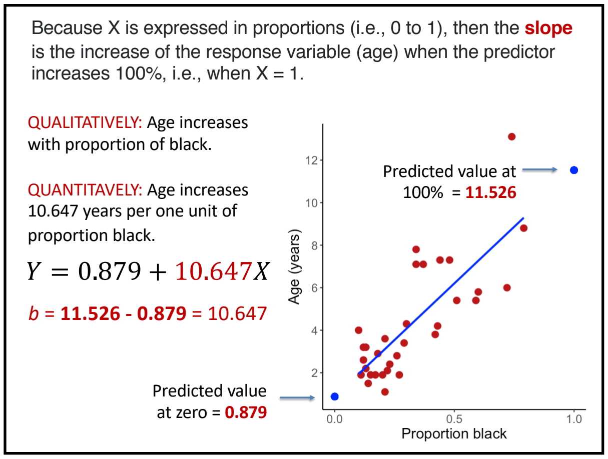
Slope b: the rate of change in y (age) as x changes (proportion black).

The slope measures the change in age of male lions per unit increase in the proportion of black.

QUALITATIVELY: Age increases with proportion of black.

QUANTITAVELY: Age increases 10.647 years per one unit of proportion black.

19



Because X is expressed in proportions (i.e., 0 to 1), then the **slope** is the increase of the response variable (age) when the predictor increases 100%, i.e., when X = 1.

QUALITATIVELY: Age increases with proportion of black.

QUANTITAVELY: Age increases 10.647 years per one unit of proportion black.

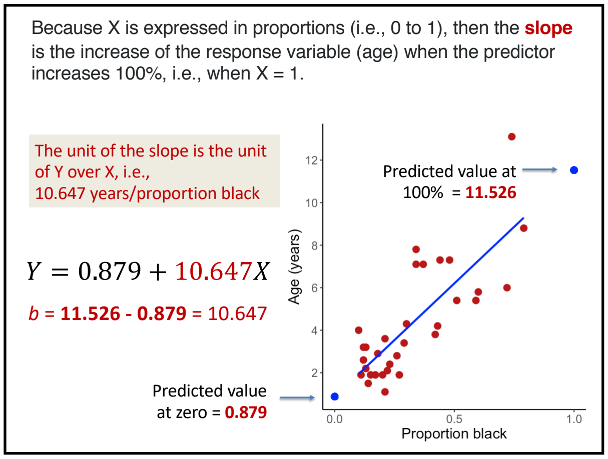
$Y = 0.879 + 10.647X$

b = **11.526 - 0.879** = 10.647

Predicted value at 100% = **11.526**

Predicted value at zero = **0.879**

20



Because X is expressed in proportions (i.e., 0 to 1), then the **slope** is the increase of the response variable (age) when the predictor increases 100%, i.e., when X = 1.

The unit of the slope is the unit of Y over X, i.e., 10.647 years/proportion black

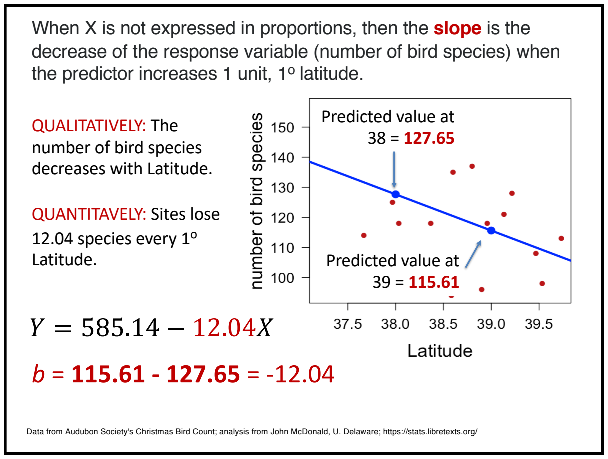
$Y = 0.879 + 10.647X$

b = **11.526 - 0.879** = 10.647

Predicted value at 100% = **11.526**

Predicted value at zero = **0.879**

21



When X is not expressed in proportions, then the **slope** is the decrease of the response variable (number of bird species) when the predictor increases 1 unit, 1° latitude.

QUALITATIVELY: The number of bird species decreases with Latitude.

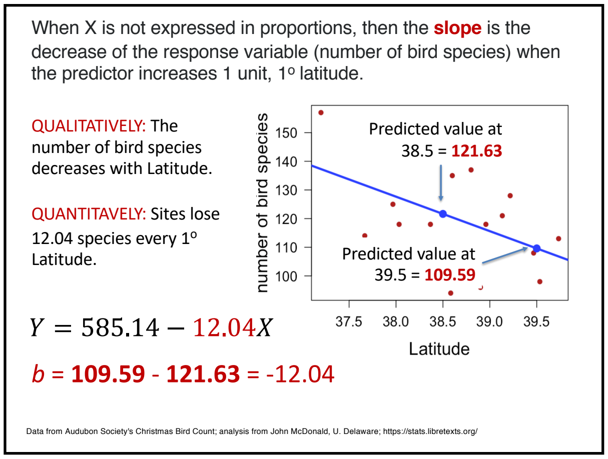
QUANTITAVELY: Sites lose 12.04 species every 1° Latitude.

$Y = 585.14 - 12.04X$

b = **115.61 - 127.65** = -12.04

Data from Audubon Society's Christmas Bird Count; analysis from John McDonald, U. Delaware; https://stats.libretexts.org/

22



When X is not expressed in proportions, then the **slope** is the decrease of the response variable (number of bird species) when the predictor increases 1 unit, 1° latitude.

QUALITATIVELY: The number of bird species decreases with Latitude.

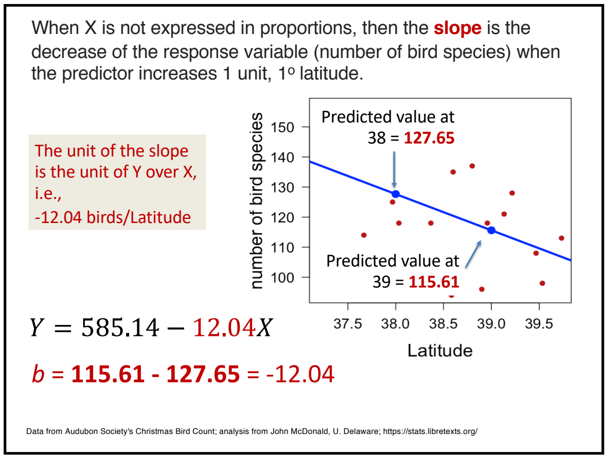
QUANTITAVELY: Sites lose 12.04 species every 1° Latitude.

$Y = 585.14 - 12.04X$

b = **109.59 - 121.63** = -12.04

Data from Audubon Society's Christmas Bird Count; analysis from John McDonald, U. Delaware; https://stats.libretexts.org/

23



When X is not expressed in proportions, then the **slope** is the decrease of the response variable (number of bird species) when the predictor increases 1 unit, 1° latitude.

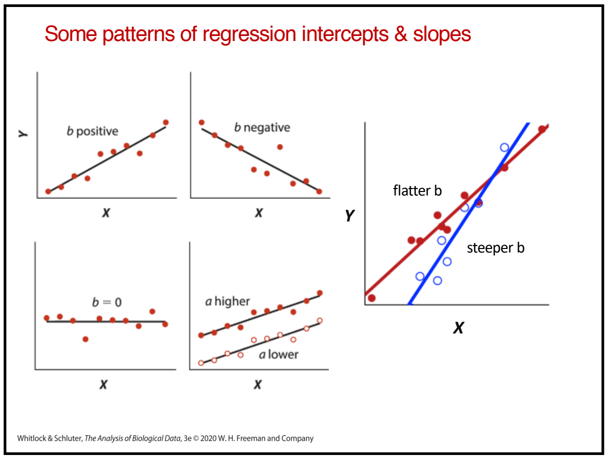
The unit of the slope is the unit of Y over X, i.e., -12.04 birds/Latitude

$Y = 585.14 - 12.04X$

b = **115.61 - 127.65** = -12.04

Data from Audubon Society's Christmas Bird Count; analysis from John McDonald, U. Delaware; https://stats.libretexts.org/

24



Some patterns of regression intercepts & slopes

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

25