

1

---

---

---

---

---

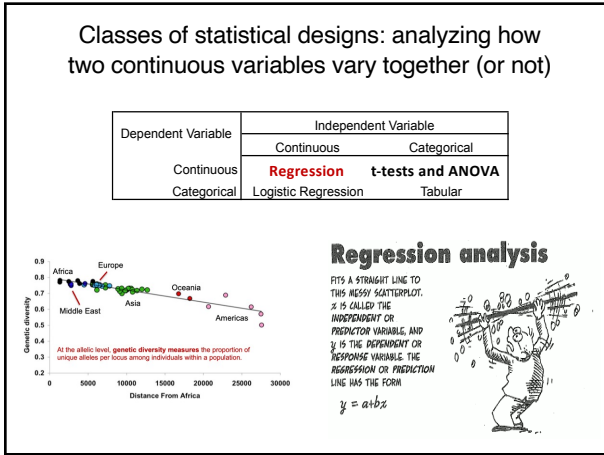
---

---

---

---

---



2

---

---

---

---

---

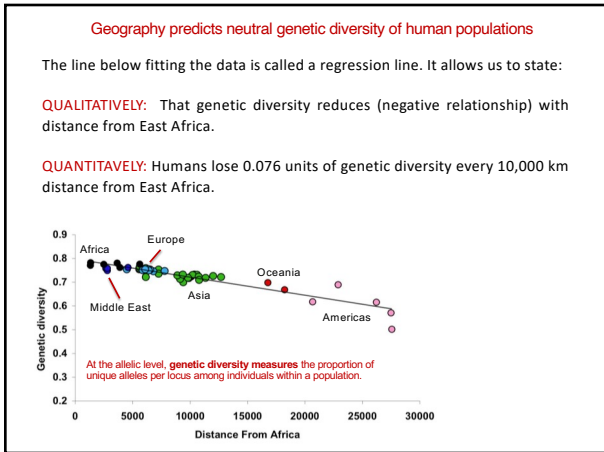
---

---

---

---

---



3

---

---

---

---

---

---

---

---

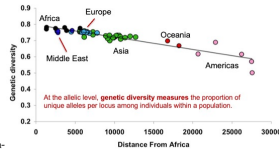
---

---

**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”.



**Why is it called “regression”?**  
<http://blog.minitab.com/blog/statistics-and-quality-data-analysis/so-why-is-it-called-regression-anyway>

---

---

---

---

---

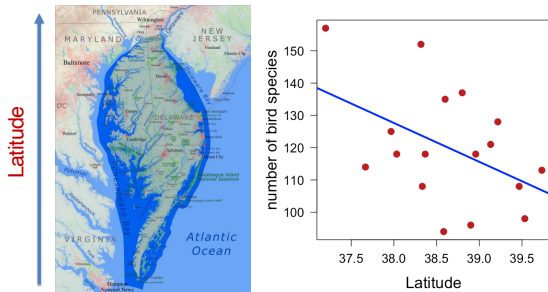
---

---

---

4

**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/>

---

---

---

---

---

---

---

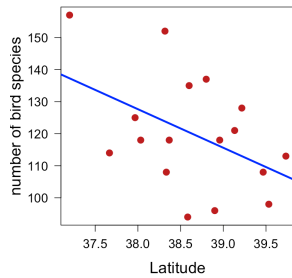
---

5

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

**QUALITATIVELY:** The number of bird species decreases with Latitude.

**QUANTITATIVELY:** Sites lose 12.04 species every 1° Latitude.



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

---

---

---

---

---

---

---

---

6

**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

---

---

---

---

---

---

---

---

---

---

7

**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
0.28	2.5	0.79	8.8
0.21	3.6	0.61	6.4

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

---

---

---

---

---

---

---

---

---

---

8

**How to fit a regression model? Some basic "jargon"**

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

---

---

---

---

---

---

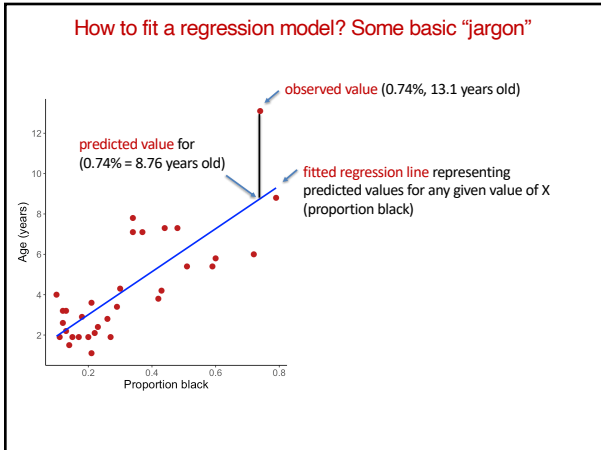
---

---

---

---

9




---

---

---

---

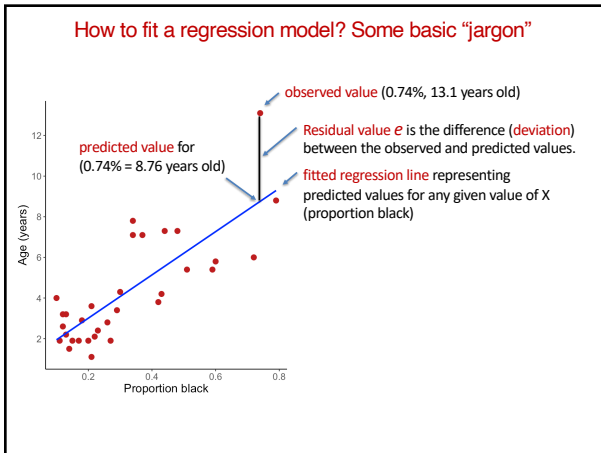
---

---

---

---

10




---

---

---

---

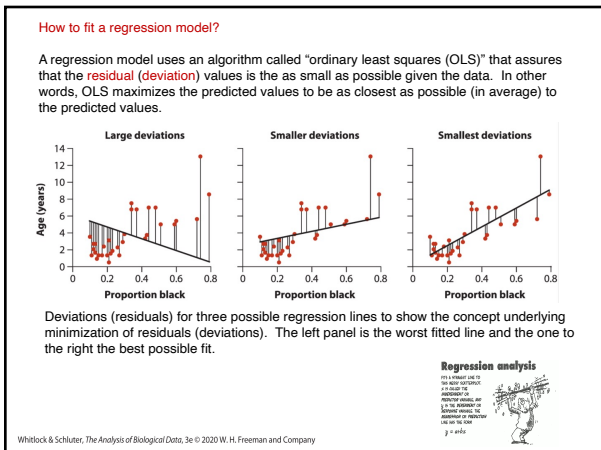
---

---

---

---

11




---

---

---

---

---

---

---

---

12

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).

---

---

---

---

---

---

---

---

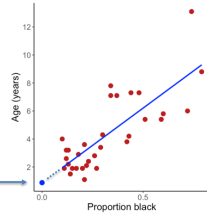
13

$$Y = a + bX \quad a = 0.879 \quad b = 10.647$$

**intercept** slope  $Y = 0.879 + 10.647X$

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

$a = 0.879$  years



Be careful when interpreting the intercept: a meaningful interpretation is only possible if  $X$  can truly be zero AND if the data include values close to zero (not the case here). For instance, the intercept could be negative in these data, but a lion cannot have a negative age.

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

---

---

---

---

---

---

---

---

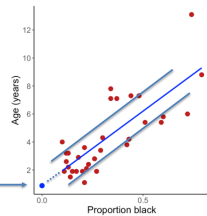
14

$$Y = a + bX \quad a = 0.879 \quad b = 10.647$$

**intercept** slope  $Y = 0.879 + 10.647X$

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

$a = 0.879$  years



The intercept is useful for prediction because it represents the addition (offset) required to correctly position the regression line so that predictions match the observed data. Different intercepts would lead to predicted values that are either too high or too low.

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

---

---

---

---

---

---

---

---

15

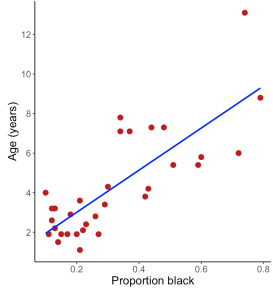
$Y = a + bX$        $a = 0.879$     $b = 10.647$   
 intercept   slope  
 $Y = 0.879 + 10.647X$

**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.

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




---

---

---

---

---

---

---

---

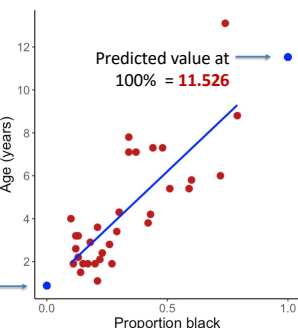
16

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.

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

$Y = 0.879 + 10.647X$   
 $b = 11.526 - 0.879 = 10.647$




---

---

---

---

---

---

---

---

17

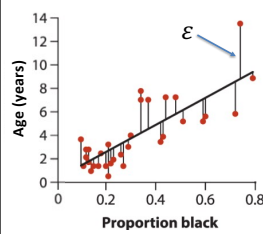
Residuals - the unexplained variation in  $Y$  (age in years) by the regression model

$Y = 0.879 + 10.647X + \epsilon$   
 $\hat{Y} = 0.879 + 10.647X$   
 $\epsilon = Y - \hat{Y}$

$\hat{Y}$  (y hat) stands for predicted values.  
 $\epsilon$  (epsilon) stands for residuals.

Residual values  $\epsilon$  are the difference (deviation) between the observed and predicted values.

Each observation in the data has a residual value.



Whitlock & Schluter, The Analysis of Biological Data, 3e © 2020 W. H. Freeman and Company  
 Sustainable trophy hunting of African lions  
 Whitman et al. (2004), Nature, 428: 175-178.

---

---

---

---

---

---

---

---

18

Residual values  $\epsilon$  are the difference (deviation) between the observed and predicted values. Predicted values  $\hat{Y}$  for each observation is on the regression line. As such, given an X value we can predict the Y value. Each observation in the data has a predicted & residual value.

	X	Y	$\hat{Y}$	$\epsilon$
1	0.21	1.4	2.114081	-2.81408129
2	0.14	1.5	2.369683	-0.86968293
3	0.11	1.9	2.050189	-0.15018934
4	0.13	2.2	2.263132	-0.06313173
5	0.12	2.6	2.126661	0.44333396
6	0.13	3.2	2.263132	0.93686827
...	...	...	...	...
28	0.37	7.1	4.818448	2.28155968
29	0.34	7.1	4.499027	2.68097318
30	0.74	13.1	8.757875	4.34212541
31	0.79	8.8	9.290251	-0.49025056
32	0.51	5.4	6.389837	-0.98983712

$\hat{Y} = 0.879 + 10.647 \times 0.51$   
 $6.31 = 0.879 + 10.647 \times 0.51$   
 $\epsilon = 5.4 - 6.31 = -0.91$   
 $5.4 = 0.879 + 10.647 \times 0.51 - 0.91$

---

---

---

---

---

---

---

---

---

---

19

### How to fit the model?

**Aim** of linear regression is to fit a straight line to data that generates (in average) the best prediction of y for any value of x.

**Predicted values** for Y are on the regression line, i.e., given an X value we can predict the Y value.

**The line** minimises the average distance between data and fitted line, i.e., the residuals.

To find the best line, we must minimise the sum of the squares of the residuals; as such we need to find model coefficients (a, b) that minimize the sum of squares of residuals:

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$


---

---

---

---

---

---

---

---

---

---

20

### How to fit the model?

To find the best line we must minimise the sum of the squares of the residuals; as such we need to find model coefficients (a & b) that minimize the sum of squares residuals:

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

There is only one such combination of a and b coefficients!!! There is a simple algorithm (method) that finds that combination: the "Ordinary Least Squares (OLS).

**Regression analysis**  
 FITS A STRAIGHT LINE TO THIS NOISY SCATTERPLOT. IT IS CALLED THE INDEPENDENT OR PREDICTOR VARIABLE AND Y IS THE DEPENDENT OR RESPONSE VARIABLE. THE REGRESSION OR PREDICTION LINE HAS THE FORM  $y = a + bx$

$Y = a + bX$

---

---

---

---

---

---

---


---

---

---

21

How to fit the model? In R



```

> (lm.fit <- lm(Age~PropBlack, data=lions))
Call:
lm(formula = Age ~ PropBlack, data = lions)

Coefficients:
(Intercept)  PropBlack
  0.879      10.647
    
```

**QUALITATIVELY:** Age increases with proportion of black.

**QUANTITATIVELY:** Age increases 10.647 years per one unit of proportion black, i.e.,  $b = 10.647$  years/proportion of black.

$Y = 0.879 + 10.647X$

---

---

---

---

---

---

---

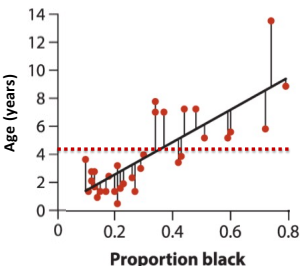
---

22

Statistical hypothesis testing in regression

**H<sub>0</sub>:** the statistical population slope  $\beta = 0$  (i.e., Y can't be predicted by X).

**H<sub>A</sub>:** the population slope  $\beta \neq 0$  (i.e., Y can be predicted by X).



As with any estimate based on sample data, slopes can differ from zero even when the true population slope is zero, simply due to sampling variation.

---

---

---

---

---

---

---

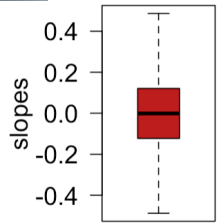
---

23

```

slopes <- c()
for (i in 1:10000){
  X <- rnorm(32)
  e <- rnorm(32)
  Y <- 0.879 + 0*X + e
  lm.fit <- lm(Y ~ X)
  slopes[i] <- lm.fit$coefficients["X"]
}
boxplot(slopes,col="firebrick",outline = FALSE,
        ylab="slopes",las = 1,cex.axis=1.3,cex.lab=1.3)
    
```

As with any estimate based on sample data, slopes can differ from zero even when the true population slope is zero, simply due to sampling variation.



---

---

---

---

---

---

---

---

24

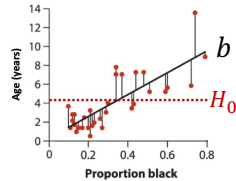
Testing whether the regression slope differs from zero:  
[1] using a t-test

$H_0$ : the statistical population slope  $\beta = 0$  (i.e., Y can't be predicted by X).

$H_A$ : the population slope  $\beta \neq 0$  (i.e., Y can be predicted by X).

The regression slope  $b$  divided by its standard error can be used to test the null hypothesis that  $\beta = 0$ . This is similar to the one-sample t-test:

$$t = \frac{b - \beta_{H_0}}{SE_b} = \frac{b - 0}{SE_b}$$




---

---

---

---

---

---

---

---

25

Testing whether the regression slope differs from zero:  
[1] using a t-test (loss of two degrees of freedom by using variance of X and Y to estimate the regression coefficients;  $df = 32-2=30$ )

```
> summary(lm(Age~PropBlack, data=lions))
Call:
lm(formula = Age ~ PropBlack, data = lions)

Residuals:
    Min       1Q   Median       3Q      Max
-2.5449 -1.1117 -0.5285  0.9635  4.3421

Coefficients:
(Intercept)  6.8798  0.5688  1.545  0.133
PropBlack   10.6471  1.5095  7.053 7.68e-08 ***
---

```

$$t = \frac{10.64}{1.51} = 7.053395$$

The t-test for the intercept is not important for the purposes of BIOL322 and simple applications of linear regressions.

$P < 0.05$ ; reject the  $H_0$  and conclude that the regression model can predict age of lions.

But can we trust its predictions? More on that later.

---

---

---

---

---

---

---

---

26

Testing whether the regression slope differs from zero:  
[2] using ANOVA (same  $H_0$  and  $H_A$ ).

```
anova(lm(Age~PropBlack, data=lions))
Analysis of Variance Table

Response: Age
Df Sum Sq Mean Sq F value Pr(>F)
PropBlack  1 138.544 138.544  49.751 7.677e-08 ***
Residuals 30  83.543   2.785
---

summary(lm(Age~PropBlack, data=lions))
Coefficients:
(Intercept)  6.8798  0.5688  1.545  0.133
PropBlack   10.6471  1.5095  7.053 7.68e-08 ***

```

$$t = \frac{10.64}{1.51} = 7.053395$$

$$F = 49.75 = t^2 = 7.053395^2 = 49.75$$

In simple regression, the t-test for the slope and the ANOVA for the regression model test the same hypothesis; in more complex models, however, ANOVA serves a different role (not covered in BIOL 322).

loss of two degrees of freedom by using variance of X and Y to estimate the regression coefficients;  $df = 32-2=30$

---

---

---

---

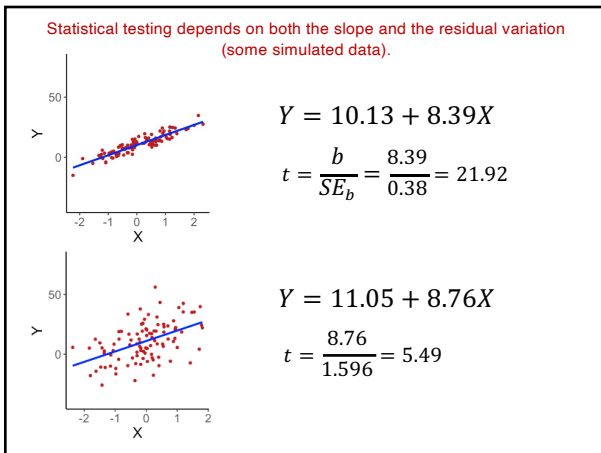
---

---

---

---

27




---

---

---

---

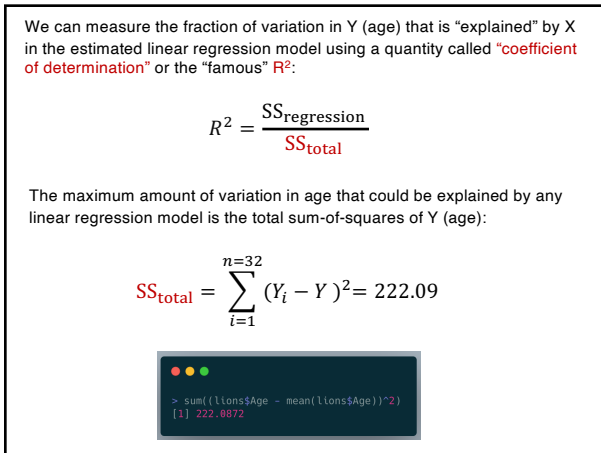
---

---

---

---

28




---

---

---

---

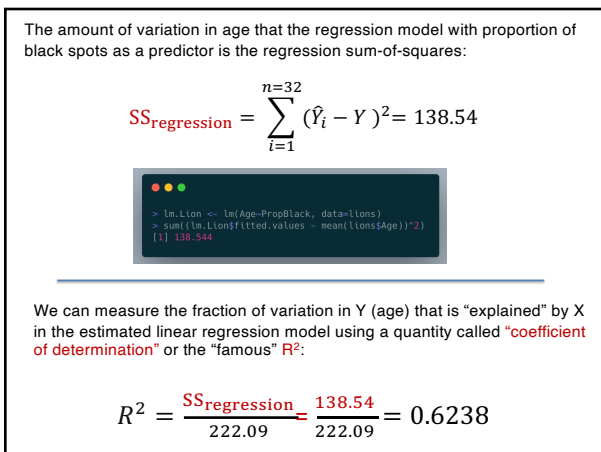
---

---

---

---

29




---

---

---

---

---

---

---

---

30

We can measure the fraction of variation in Y (age) that is "explained" by X in the estimated linear regression model using a quantity called "coefficient of determination" or the "famous"  $R^2$ :

$$R^2 = \frac{SS_{\text{regression}}}{222.09} = \frac{138.54}{222.09} = 0.6238$$

We state then that the regression model explains 62.38% of the total variation in age.

```

> summary(lm(Age~PropBlack, data=lions))
Coefficients:
(Intercept)  0.6798    0.5688    1.545    0.133
PropBlack    10.6471    1.5955    7.653    7.60e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared:  0.6238,    Adjusted R-squared:  0.6113
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
    
```

$R^2 = 0.6238$

Adjusted  $R^2$  is a more complex measure, and we cover it in BIOL 422.

---

---

---

---

---

---

---

---

---

---

31

Residuals, not just the slope, influence how well a regression model explains variation in Y (some simulated data).

$Y = 10.13 + 8.39X$   
 $t = \frac{b}{SE_b} = \frac{8.39}{0.38} = 21.9$   
 $R^2 = 0.8289 = 83.89\%$

$Y = 11.05 + 8.76X$   
 $t = \frac{8.76}{1.596} = 5.49$   
 $R^2 = 0.2275 = 22.75\%$

---

---

---

---

---

---

---

---

---

---

32

The last sum-of-squares involved in a regression:

$$SS_{\text{residuals}} = \sum_{i=1}^{n=32} e_i^2 = 83.54$$

```

> lm.Lion <- lm(Age~PropBlack, data=lions)
> sum((lm.Lion$residuals)^2)
[1] 83.54321
    
```

---

---

---

---

---

---

---

---

---

---

33

All the **sum-of-squares** involved:

$$SS_{\text{regression}} = \sum_{i=1}^{n=32} (\hat{Y}_i - Y)^2 = 138.54 \quad SS_{\text{total}} = \sum_{i=1}^{n=32} (Y_i - Y)^2 = 222.09$$

$$SS_{\text{residuals}} = \sum_{i=1}^{n=32} e_i^2 = 83.54$$

$$SS_{\text{total}} = SS_{\text{regression}} + SS_{\text{residuals}}$$

$$222.09 = 138.544 + 83.544$$


---

---

---

---

---

---

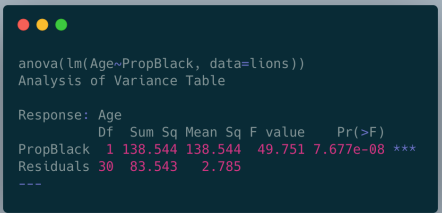
---

---

34

The three **sum-of-squares** involved in a regression and its relation to the F statistic and F-distribution:

$$F = \frac{SS_{\text{regression}}/df_{\text{regression}}}{SS_{\text{residual}}/df_{\text{residual}}}$$

$$\frac{SS_{\text{regression}}/1}{SS_{\text{residual}}/(n-2)} = \frac{138.54/1}{83.54/30} = 49.75$$


```

anova(lm(Age~PropBlack, data=lions))
Analysis of Variance Table

Response: Age
      Df Sum Sq Mean Sq F value    Pr(>F)
PropBlack  1 138.544 138.544  49.751 7.677e-08 ***
Residuals 30  83.543   2.785
---
    
```

---

---

---

---

---

---

---

---

35




---

---

---

---

---

---

---

---

36

**Using regressions to make predictions**

(regression of Y on X does not always imply dependency)  
**SPURIOUS CORRELATION**

“Predictive capacity without explanatory capacity is worthless. Mere clairvoyance, irrespective of its sharpness, does not itself have scientific standing. Only predictive capacity that arises out of having coherent and communicable explanations has scientific standing. The power to predict is subsidiary to the power to explain. Explanation without prediction is sufficient, but prediction without explanation is of no consequence from a scientific standpoint.”

—Harvey Leibenstein (1966), in “Beyond Economic Man”.

As George E. P. Box said: “All models are wrong, but some are useful”

---

---

---

---

---

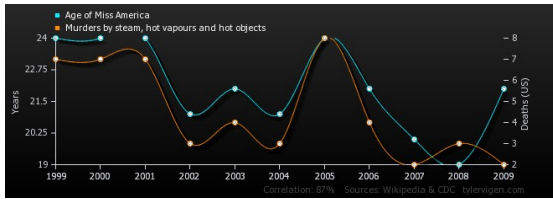
---

---

---

37

Regression of Y on X does not always imply dependency  
**SPURIOUS CORRELATION: correlation between two variables having no causal relation.**



The Regression of Divorce rate in Main on per capita consumption of margarine (US) is  $R^2 = 0.985$

<https://tylervigen.com/old-version.html>

---

---

---

---

---

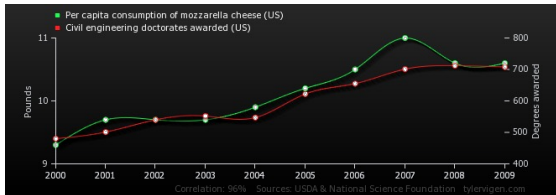
---

---

---

38

Regression of Y on X does not always imply dependency  
**SPURIOUS CORRELATION: correlation between two variables having no causal relation.**



The Regression of Civil engineering doctorates (US) on per capita consumption of mozzarella cheese is  $R^2 = 0.919$

<https://tylervigen.com/old-version.html>

---

---

---

---

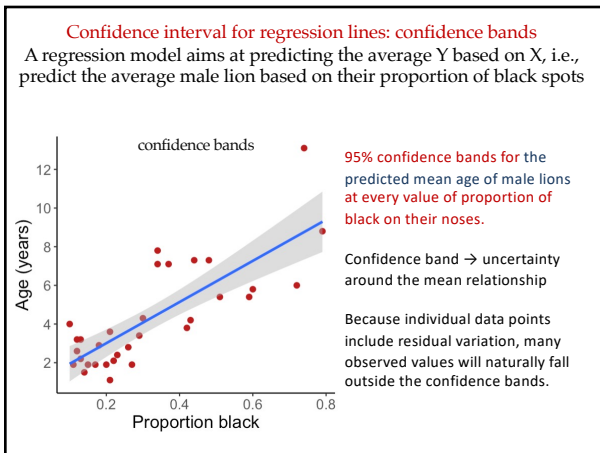
---

---

---

---

39




---

---

---

---

---

---

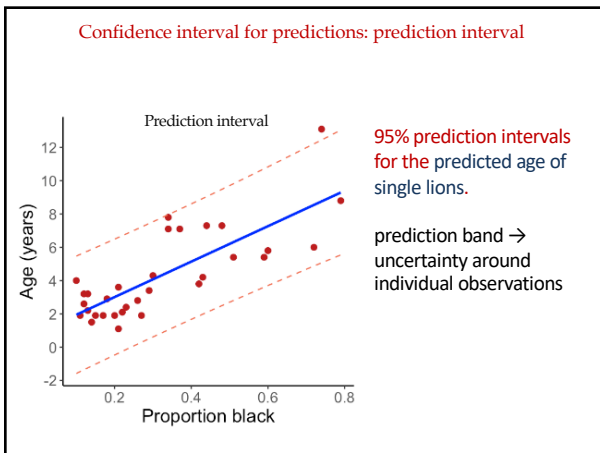
---

---

---

---

40




---

---

---

---

---

---

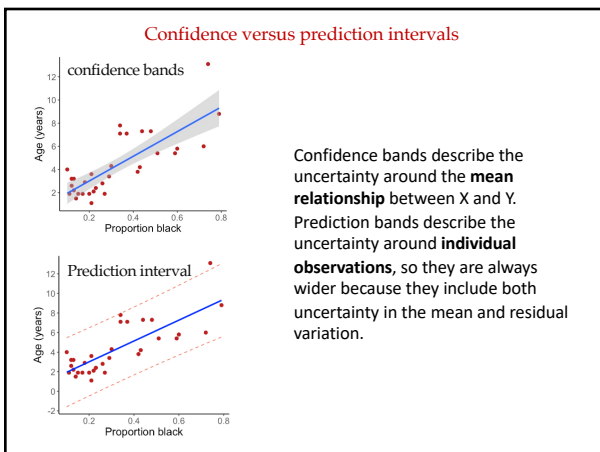
---

---

---

---

41




---

---

---

---

---

---

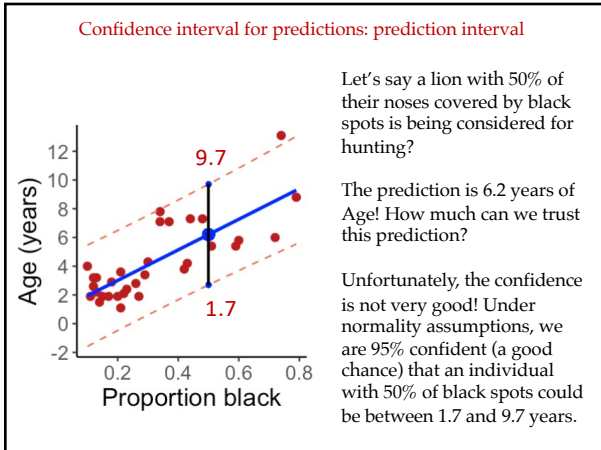
---

---

---

---

42




---

---

---

---

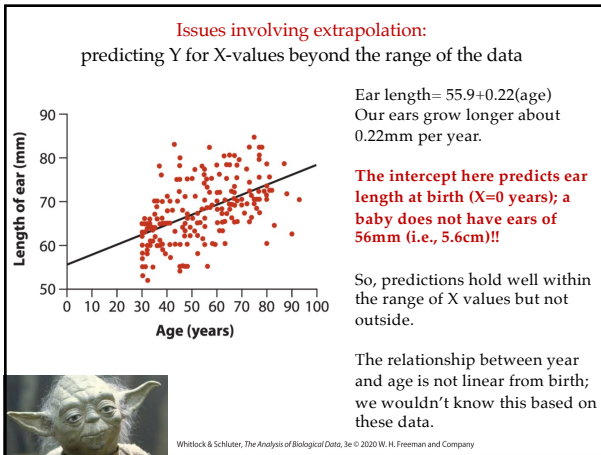
---

---

---

---

43




---

---

---

---

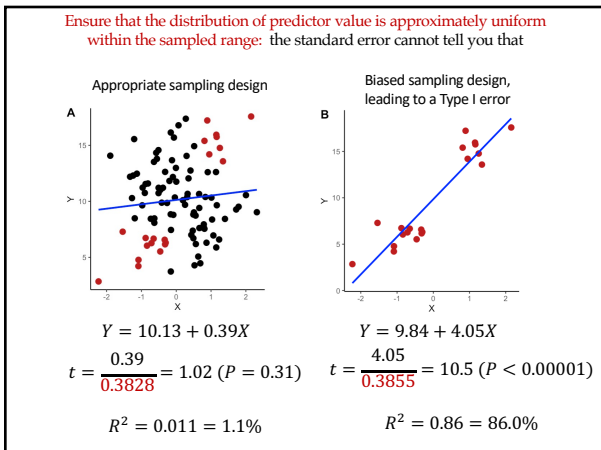
---

---

---

---

44




---

---

---

---

---

---

---

---

45

assumptions coming next



---

---

---

---

---

---

---

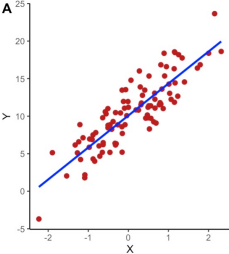
---

46

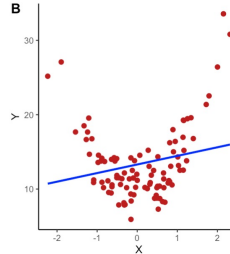
Assumptions of regression models: [1] linearity

Appropriate data for a linear model      Non-appropriate data for a linear model

**A**



**B**



It is critical to graph the data

---

---

---

---

---

---

---

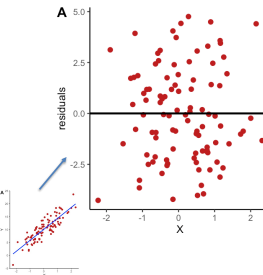
---

47

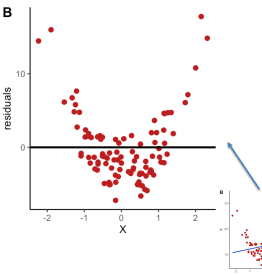
Assumptions of regression models: [1] linearity

Appropriate data for a linear model      Non-appropriate data for a linear model

**A**



**B**



Plotting the residuals against predictor values is critical in assessing whether a linear model is appropriate. The horizontal line is the average of residuals (which is always zero as a result of the fitting method). If variance is greater in different parts of the line, this indicates lack of linearity or heteroscedasticity (more on that in a few slides).

---

---

---

---

---

---

---

---

48

**Assumptions of regression models: [2] all observations have similar influences on the regression model**

**Francis Anscombe's quartets:** it comprises four data sets that have nearly identical simple descriptive statistics and regression models. Yet, they have very different distributions and appear very different when graphed. These data demonstrate both the importance of graphing data before analyzing it and the influence of influential observations (outliers).

All Quartets have the same regression model and  $R^2$ :

$$Y = -1.0 + 1.33X$$

$$R^2 = 0.63 = 63\%$$

Quartet 1 is the only appropriate in the sense that all observations have the same influence on the model, i.e., removal of one observation won't affect the model much. There are different methods to estimate the influence of each observation on the model (advanced level).

See also [https://en.wikipedia.org/wiki/Anscombe%27s\\_quartet](https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

---

---

---

---

---

---

---

---

---

---

49

**Assumptions of regression models: [3] residual variation is normally distributed**

**remember:** A regression model aims at predicting the average Y based on X, i.e., predict the average Y based on X.

**Normality assumption:** At each value of X, there is a normally distributed population of Y-values with the mean on the true regression line.

One can estimate the model even if residuals are not normally distributed, but one cannot generalize the model to predict other observations in the statistical population or make inferences (e.g., p-value, confidence intervals, t-tests, ANOVAs).

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

---

---

---

---

---

---

---

---

---

---

50

**Assumptions of regression models: [4] residual variation is homoscedastic (constant across the range of X values)**

**Homoscedasticity assumption:** At each value of X, there is a normally distributed population of Y-values with the mean on the true regression line. The variance of the Y-values is assumed to be the same for every value of X.

One can estimate the model even if residuals are not homoscedastic, but one cannot generalize the model to predict other observations in the statistical population or make inferences (e.g., p-value, confidence intervals, t-tests, ANOVAs).

---

---

---

---

---

---

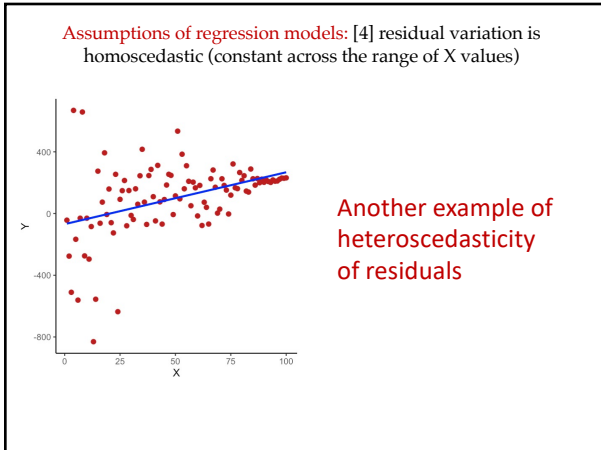
---

---

---

---

51




---

---

---

---

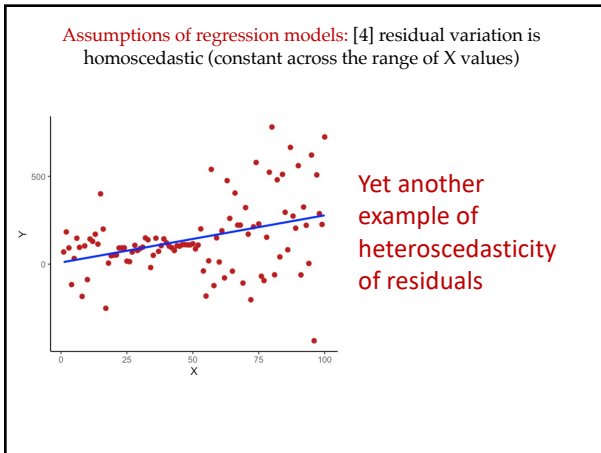
---

---

---

---

52




---

---

---

---

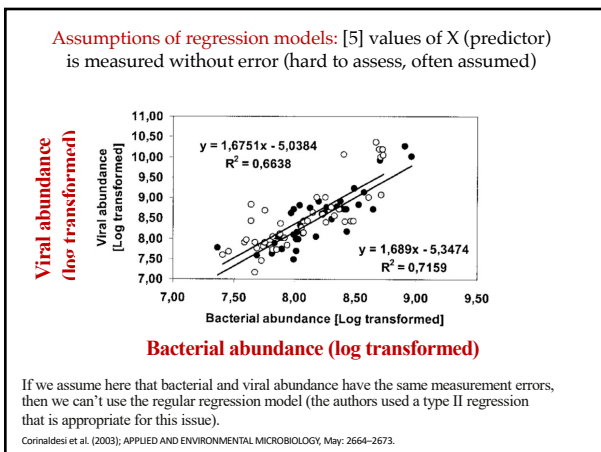
---

---

---

---

53




---

---

---

---

---

---

---

---

54



Assumptions of regression models: [5] values of X (predictor) is measured without error (hard to assess, often assumed)

$Y = 0.879 + 1.300X$  True population model

```

slopes <- c()
slopes.error <- c()
for (i in 1:10000){
  X <- rnorm(100)
  e <- rnorm(100)
  Y <- 0.879 + 1.3*X + e
  lm.fit <- lm(Y ~ X)
  slopes[i] <- lm.fit$coefficients["X*"]
  X.error <- rnorm(100,X,sd=1)
  lm.fit <- lm(Y ~ X.error)
  slopes.error[i] <- lm.fit$coefficients["X.error"]
}
boxplot(slopes,slopes.error,col="firebrick",outline = FALSE,
        ylab="slopes",las = 1,cex.axis=1.3,cex.lab=1.3)
    
```

---

---

---

---

---

---

---

---

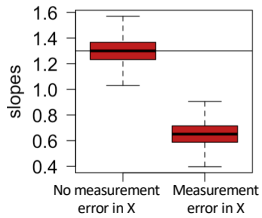
---

---

58

Assumptions of regression models: [5] values of X (predictor) is measured without error (hard to assess, often assumed)

$Y = 0.879 + 1.300X$  True population model



```

slopes <- c()
slopes.error <- c()
for (i in 1:10000){
  X <- rnorm(100)
  e <- rnorm(100)
  Y <- 0.879 + 1.3*X + e
  lm.fit <- lm(Y ~ X)
  slopes[i] <- lm.fit$coefficients["X"]
  X.error <- rnorm(100,X,sd=1)
  lm.fit <- lm(Y ~ X.error)
  slopes.error[i] <- lm.fit$coefficients["X.error"]
}
boxplot(slopes,slopes.error,col="firebrick",outline = FALSE,
        ylab="slopes",las = 1,cex.axis=1.3,cex.lab=1.3)
    
```

---

---

---

---

---

---

---

---

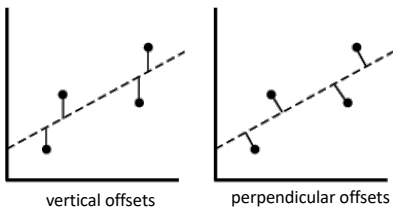
---

---

59

Assumptions of regression models: [5] values of X (predictor) is measured without error (hard to assess, often assumed)

One approach to this problem is the so-called Type II regression models (not covered in BIOL322 in details)



Residuals for Type I regression  
Error in Y but not in X

Residuals for Type II regression  
Error in both Y and X

---

---

---

---

---

---

---

---

---

---

60

**Assumptions of regression models:** [5] values of X (predictor) is measured without error (hard to assess, often assumed)

$Y = 0.879 + 1.300X$  True population model

One approach to this problem is the so-called **Type II regression models** (not covered in BIOL322)

Type II regression is not biased but greater standard error (sampling variation): no “free lunch”. This is obvious because both X and Y have errors.

No measurement error in X (Type I regression)    Measurement error in X (Type I regression)    Measurement error in X (Type II regression)

---

---

---

---

---

---

---

---

61

**Assumptions of regression models:** [6] residuals are independent: this is the assumption in which data are sampled randomly

When residuals are non independent, one should be careful about making inferences (e.g., p-value, confidence intervals, t-tests, ANOVAs); more of this issue in advanced BIOL422.

---

---

---

---

---

---

---

---

62