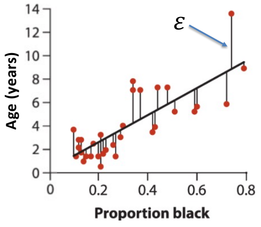
Residuals - the unexplained variation in Y (age in years) by the regression model

$$Y = 0.879 + 10.647X + \varepsilon$$

⇩

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

ε = Y-$\hat{Y}$



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

ε (epsilon) stands for residuals.

Residual values ε are the difference (deviation) between the observed and predicted values.

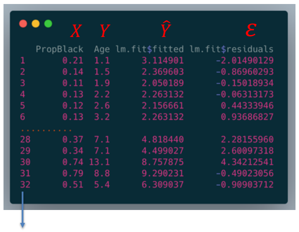
Each observation in the data has a residual value.

Sustainable trophy hunting of African lions
Whitman et al. (2004), Nature, 428: 175-178.

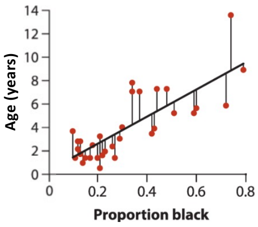
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

1

Residual values ε 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.



```
     PropBlack  Age lm.fit$fitted lm.fit$residuals
1         0.21  1.1      3.114901      -2.01490129
2         0.14  1.5      2.369603      -0.86960293
3         0.11  1.9      2.050189      -0.15018934
4         0.13  2.2      2.263132      -0.06313173
5         0.12  2.6      2.156661       0.44333946
6         0.13  3.2      2.263132       0.93686827
..........
28        0.37  7.1      4.818440       2.28155960
29        0.34  7.1      4.499027       2.60097318
30        0.74 13.1      8.757875       4.34212541
31        0.79  8.8      9.290231      -0.49023056
32        0.51  5.4      6.309037      -0.90903712
```



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

2

How to fit the model?

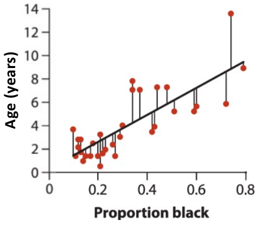
Aim of linear regression is to fit a straight line to data that generates (in average) the best prediction of y for any value of x.

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:

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



3

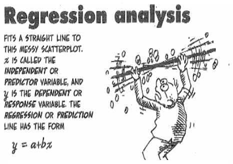
How to fit the model?

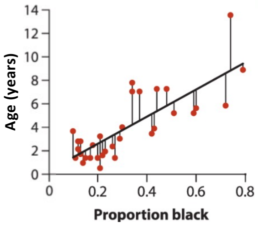
To find the best line we must minimise the sum of the squares of the residuals; as such we need to find model coefficients (a & b) that minimize the sum of of squares residuals:

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

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).

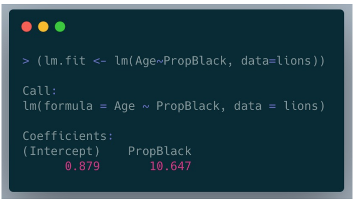
$$Y = a + bX$$


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



4

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.

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

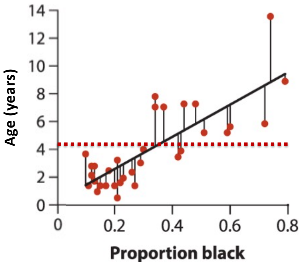
$$Y = 0.879 + 10.647X$$

5

Statistical hypothesis testing in regression

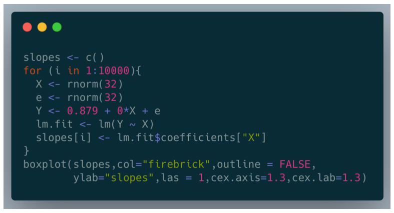
H$_0$: the statistical population slope $\beta = 0$ (i.e., Y can’t be predicted by X).

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



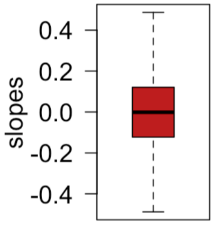
As for any other estimate (i.e., based on sample data), slopes can differ from 0 even if they came from a statistical population where the regression slope is zero.

6



```
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 for any other estimate (i.e., based on sample data), slopes can differ from 0 even if they came from a statistical population where the regression slope is zero.



7

Testing whether the regression slope differs from zero:

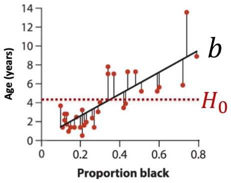
[1] using a t-test

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:

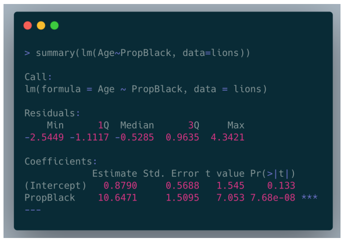
$$t = \frac{b - \beta_{H_0}}{SE_b} = \frac{b - 0}{SE_b}$$



8

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:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.8790     0.5688   1.545    0.133    
PropBlack    10.6471     1.5095   7.053 7.68e-08 ***
---
```

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

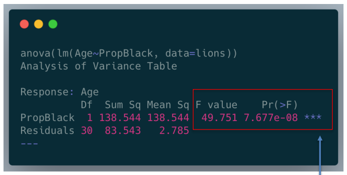
The t-test for the intercept is not important for the purposes of BIOL322 and simple applications of linear regressions.

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

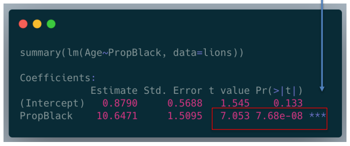
But can we trust its predictions? More on that later.

9

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:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.8790     0.5688   1.545    0.133    
PropBlack    10.6471     1.5095   7.053 7.68e-08 ***
```

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

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

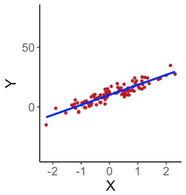
In simple regression, the t-test for slopes and ANOVA for the regression model are the same thing; in more complex models, ANOVA plays a different role (not covered in BIOL322).

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

10

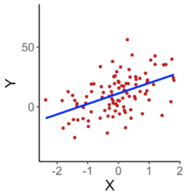
Residuals (not the slope) influence its error and statistical testing

(some simulated data)



$$Y = 10.13 + 8.39X$$

$$t = \frac{b}{SE_b} = \frac{8.39}{0.38} = 21.92$$



$$Y = 11.05 + 8.76X$$

$$t = \frac{8.76}{1.596} = 5.49$$

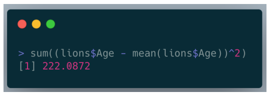
11

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:

$$R^2 = \frac{\text{SS}_{\text{regression}}}{\text{SS}_{\text{total}}}$$

The maximum amount of variation in age that could be explained by any linear regression model is the total sum-of-squares of Y (age):

$$\text{SS}_{\text{total}} = \sum_{i=1}^{n=32} (Y_i - Y\ )^2 = 222.09$$

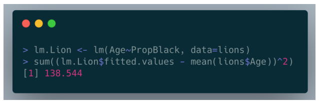


```
> sum((lions$Age - mean(lions$Age))^2)
[1] 222.0872
```

12

The amount of variation in age that the regression model with proportion of black spots as a predictor is the regression sum-of-squares:

$$\mathrm{SS}_{\mathrm{regression}} = \sum_{i=1}^{n=32} (\hat{Y}_i - Y\ )^2 = 138.54$$



```
> lm.Lion <- lm(Age~PropBlack, data=lions)
> sum((lm.Lion$fitted.values - mean(lions$Age))^2)
[1] 138.544
```

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:

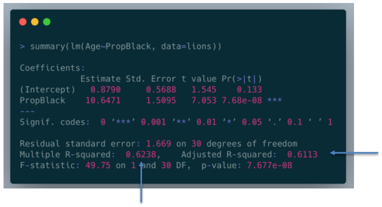
$$R^2 = \frac{\mathrm{SS}_{\mathrm{regression}}}{222.09} = \frac{138.54}{222.09} = 0.6238$$

13

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:

$$R^2 = \frac{\mathrm{SS}_{\mathrm{regression}}}{222.09} = \frac{138.54}{222.09} = 0.6238$$

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



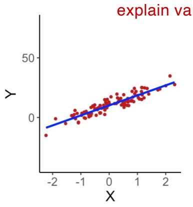
```
> summary(lm(Age~PropBlack, data=lions))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8790     0.5688   1.545    0.133
PropBlack    10.6471     1.5095   7.053 7.68e-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
```

The adjusted-R^2 is a more complex estimator and we leave it for BIOL422.

$R^2 = 0.6238$

14

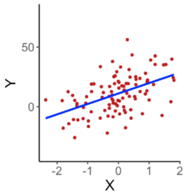
Residuals (not the slope) influence the ability of a regression model to explain variation in Y (some simulated data)



$$Y = 10.13 + 8.39X$$

$$t = \frac{b}{SE_b} = \frac{8.39}{0.38} = 21.9$$

$$\boldsymbol{R^2 = 0.8289 = 83.89\%}$$



$$Y = 11.05 + 8.76X$$

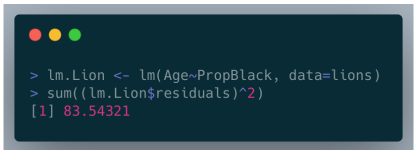
$$t = \frac{8.76}{1.596} = 5.49$$

$$\boldsymbol{R^2 = 0.2275 = 22.75\%}$$

15

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

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



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

16

All the sum-of-squares involved:

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

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

$$\text{SS}_{\text{total}} = \text{SS}_{\text{regression}} + \text{SS}_{\text{residuals}}$$

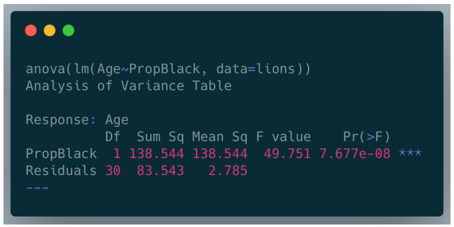
$$222.09 = 138.544 + 83.544$$

17

All the sum-of-squares involved in a regression and its relation to F:

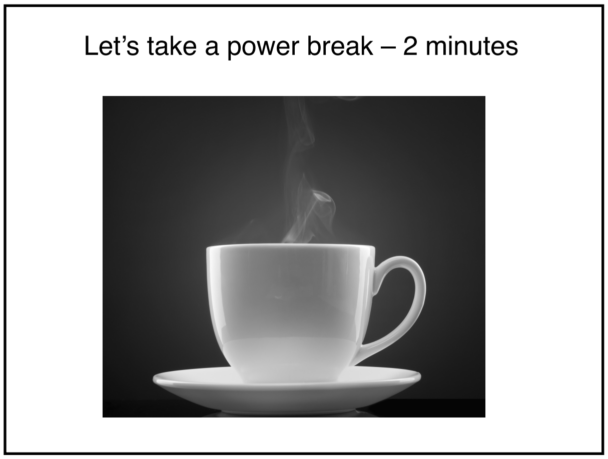
$$F = \frac{\text{SS}_{\text{regression}}/\text{df}_{\text{regression}}}{\text{SS}_{\text{residual}}/\text{df}_{\text{residual}}} =$$

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



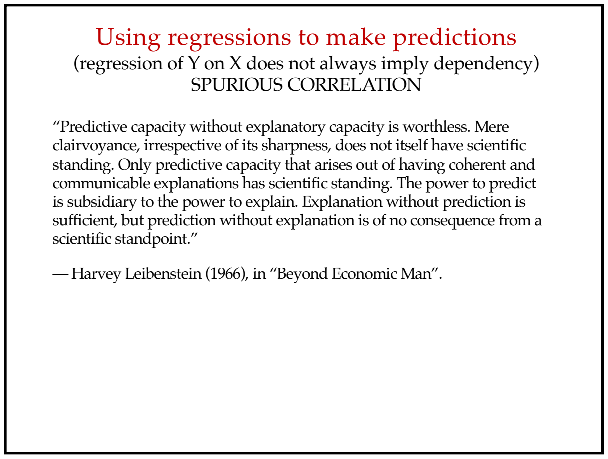
```
anova(lm(Age~PropBlack, data=lions))
Analysis of Variance Table

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
---
```

18



Let’s take a power break – 2 minutes

19



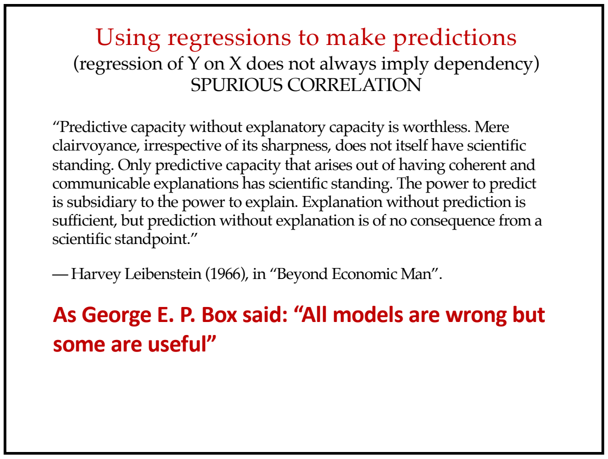
Using regressions to make predictions

(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”.

20



Using regressions to make predictions

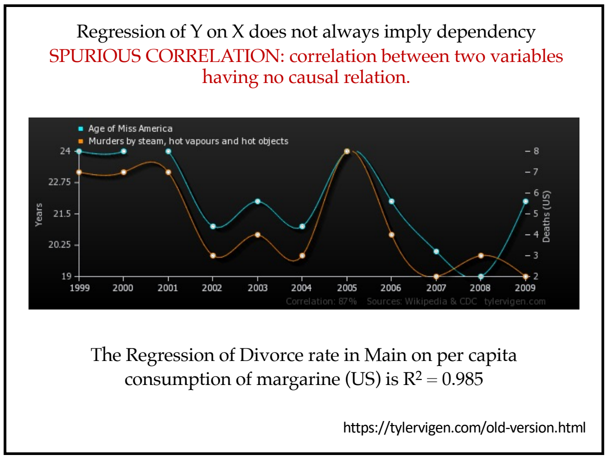
(regression of Y on X does not always imply dependency)
SPURIOUS CORRELATION

“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”

21



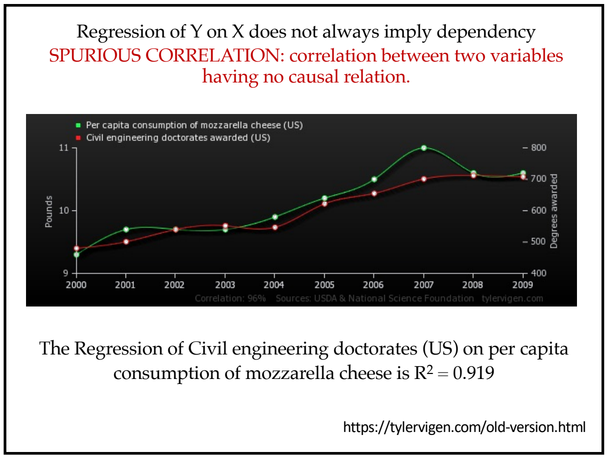
Regression of Y on X does not always imply dependency

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

22



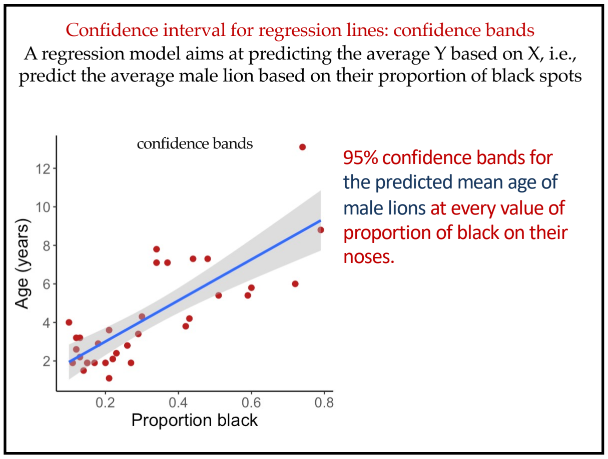
Regression of Y on X does not always imply dependency

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

23



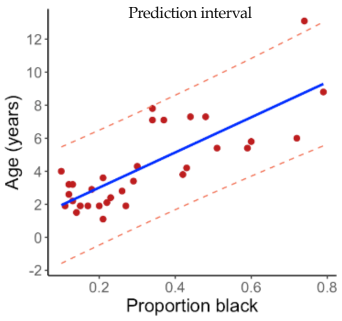
Confidence interval for regression lines: confidence bands

A regression model aims at predicting the average Y based on X, i.e., predict the average male lion based on their proportion of black spots

95% confidence bands for the predicted mean age of male lions at every value of proportion of black on their noses.

24

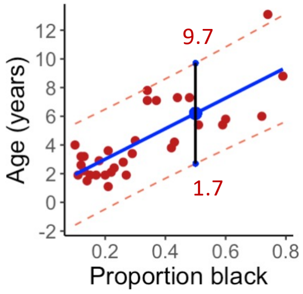
Confidence interval for predictions: prediction interval



95% prediction intervals for the predicted age of single lions.

25

Confidence interval for predictions: prediction interval



Let's say a lion with 50% of their noses covered by black spots is being considered for hunting?

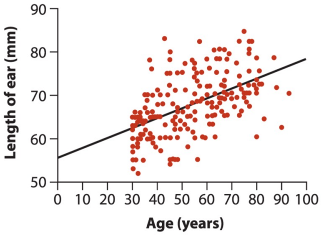
The prediction is 6.2 years of Age! How much can we trust this prediction?

Unfortunately, the confidence is not very good! Under normality assumptions, we are 95% confident (a good chance) that an individual with 50% of black spots could be between 1.7 and 9.7 years.

26

Issues involving extrapolation:

predicting Y for X-values beyond the range of the data

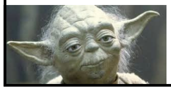


Ear length= 55.9+0.22(age)
Our ears grow longer about 0.22mm per year.

The intercept predicts ear length at birth (X=0 years); a baby does not have ears of 5.6cm!!

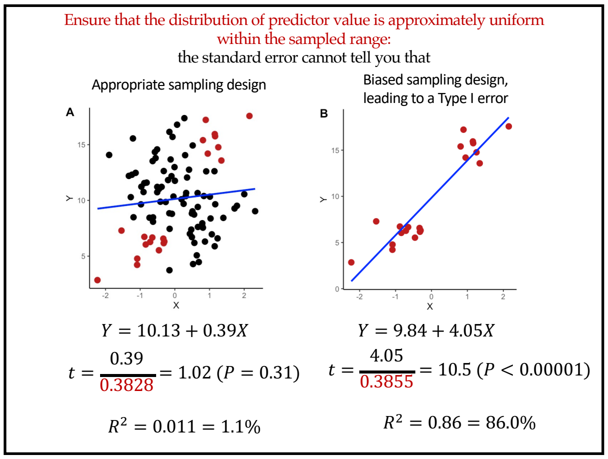
So predictions hold well within the range of X values but not outside.

The relationship between year and age is not linear from birth; we wouldn't know this based on these data.



Whitlock & Schl

27



Ensure that the distribution of predictor value is approximately uniform within the sampled range:
the standard error cannot tell you that

Appropriate sampling design

Biased sampling design, leading to a Type I error

$Y = 10.13 + 0.39X$

$t = \dfrac{0.39}{0.3828} = 1.02\ (P = 0.31)$

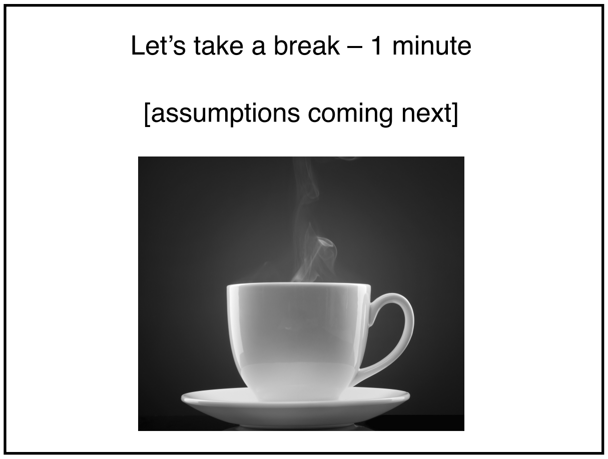
$R^2 = 0.011 = 1.1\%$

$Y = 9.84 + 4.05X$

$t = \dfrac{4.05}{0.3855} = 10.5\ (P < 0.00001)$

$R^2 = 0.86 = 86.0\%$

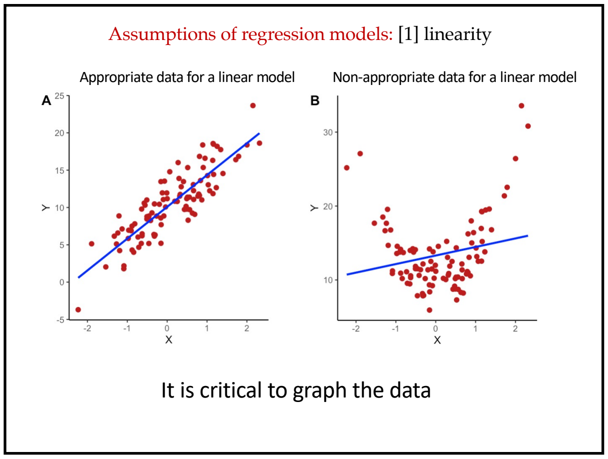
28



Let's take a break – 1 minute

[assumptions coming next]

29



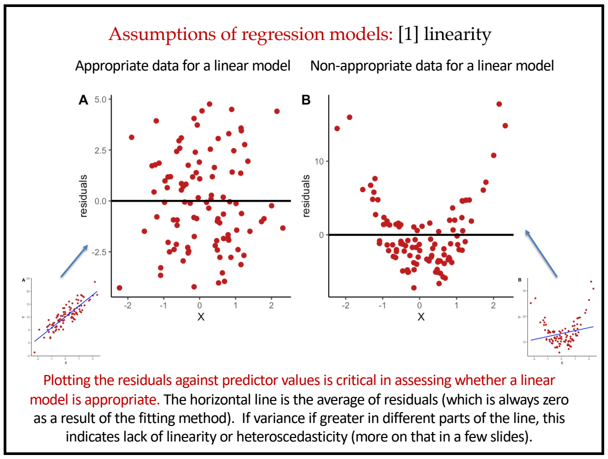
Assumptions of regression models: [1] linearity

Appropriate data for a linear model

Non-appropriate data for a linear model

It is critical to graph the data

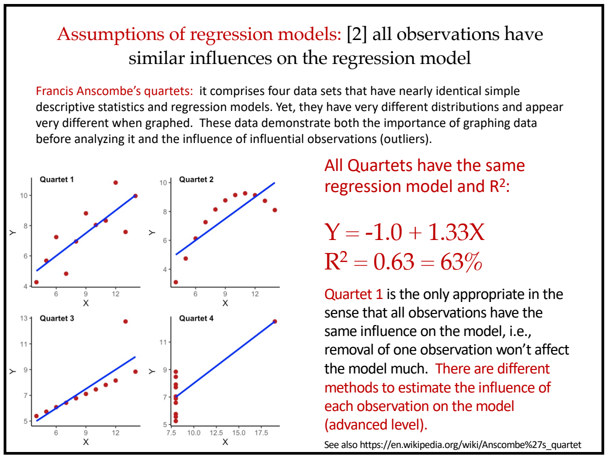
30



Assumptions of regression models: [1] linearity

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 if greater in different parts of the line, this indicates lack of linearity or heteroscedasticity (more on that in a few slides).

31



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:

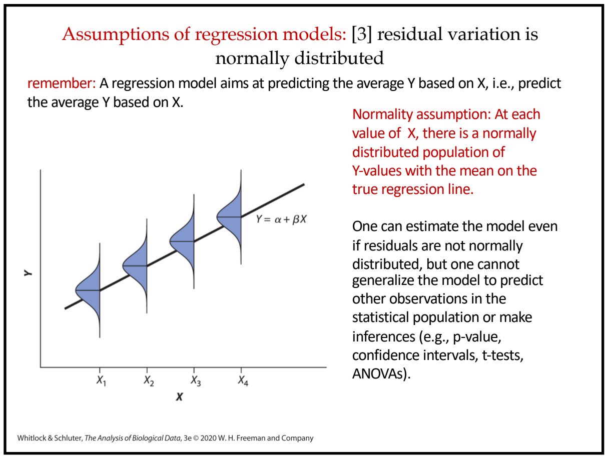
$Y = -1.0 + 1.33X$

$R^2 = 0.63 = 63\%$

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

32



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.

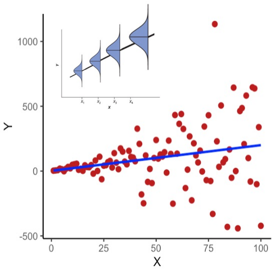
Normality assumption: At each value of X, there is a normally distributed population of Y-values with the mean on the true regression line.

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

33

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

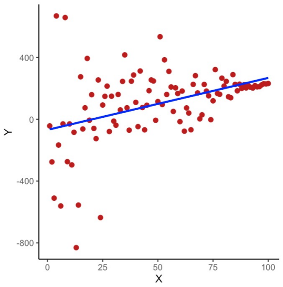


Heteroscedasticity assumption: At each value of X, there is a normally distributed population of Y-values with the mean on the true regression line. The variance of the Y-values is assumed to be the same for every value of X.

One can estimate the model even if residuals are not heteroscedastic, 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).

34

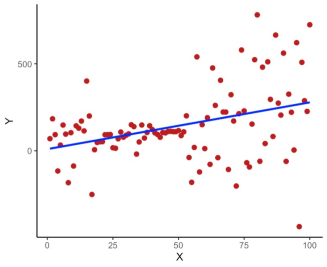
Assumptions of regression models: [4] residual variation is homoscedastic (constant across the range of X values)



Another example of heteroscedasticity of residuals

35

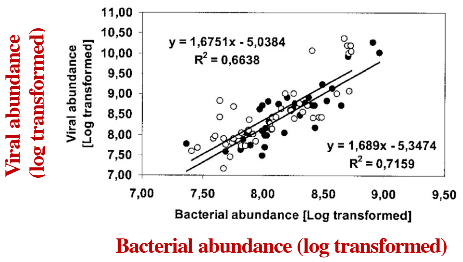
Assumptions of regression models: [4] residual variation is homoscedastic (constant across the range of X values)



Yet another example of heteroscedasticity of residuals

36

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



If we assume here that bacterial and viral abundance have the same measurement errors, then we can't use the regular regression model (the authors used a type II regression that is appropriate for this issue).

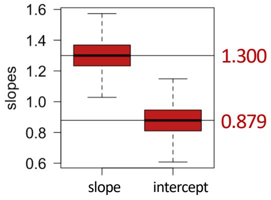
Corinaldesi et al. (2003); APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May: 2664–2673.

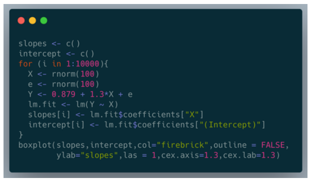
37

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

But first we need to understand (revisit) that the regression model based on samples are an unbiased estimate of the true intercepts and slopes. Let's assume the following population regression model:

$$Y = 0.879 + 1.300X$$



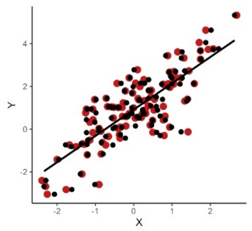


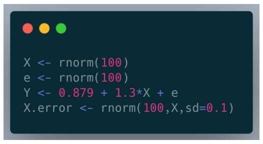
```
slopes <- c()
intercept <- c()
for (i in 1:10000){
  X <- rnorm(100)
  e <- rnorm(100)
  Y <- 0.879 + 1.3*X + e
  lm.fit <- lm(Y ~ X)
  slopes[i] <- lm.fit$coefficients["X"]
  intercept[i] <- lm.fit$coefficients["(Intercept)"]
}
boxplot(slopes,intercept,col="firebrick",outline = FALSE,
        ylab="slopes",las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

Sampling variation in estimates

38

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





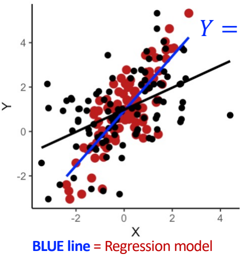
```
X <- rnorm(100)
e <- rnorm(100)
Y <- 0.879 + 1.3*X + e
X.error <- rnorm(100,X,sd=0.1)
```

Red dots are X values "measured" without error, whereas the smaller black does are X values "measured" with error.

In this case there is little consequence because the error is small (0.1).

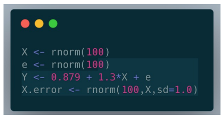
39

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



$Y = 0.929 + 1.23X$ without error in X

$Y = 0.977 + 0.498X$ with error in X



```r
X <- rnorm(100)
e <- rnorm(100)
Y <- 0.879 + 1.3*X + e
X.error <- rnorm(100,X,sd=1.0)
```

BLUE line = Regression model without error in X.

BLACK line = Regression model with error in X.

ERROR IN X REDUCES SLOPES.

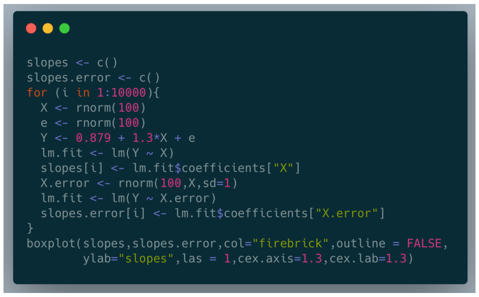
Red dots are X values “measured” without error, whereas the smaller black does are X values “measured” with error.

The consequence here is much bigger for estimating the regression model because the error is large (1.0).

40

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

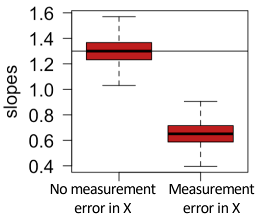


```r
slopes <- c()
slopes.error <- c()
for (i in 1:10000){
  X <- rnorm(100)
  e <- rnorm(100)
  Y <- 0.879 + 1.3*X + e
  lm.fit <- lm(Y ~ X)
  slopes[i] <- lm.fit$coefficients["X"]
  X.error <- rnorm(100,X,sd=1)
  lm.fit <- lm(Y ~ X.error)
  slopes.error[i] <- lm.fit$coefficients["X.error"]
}
boxplot(slopes,slopes.error,col="firebrick",outline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

41

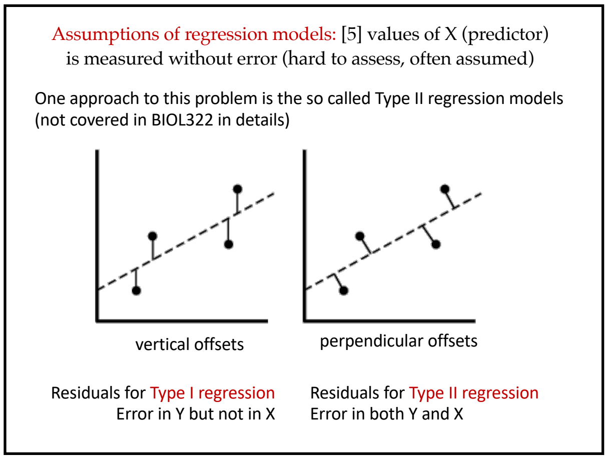
Assumptions of regression models: [5] values of X (predictor) is measured without error (hard to assess, often assumed)

$Y = 0.879 + 1.300X$ True population model



```r
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)
```

42



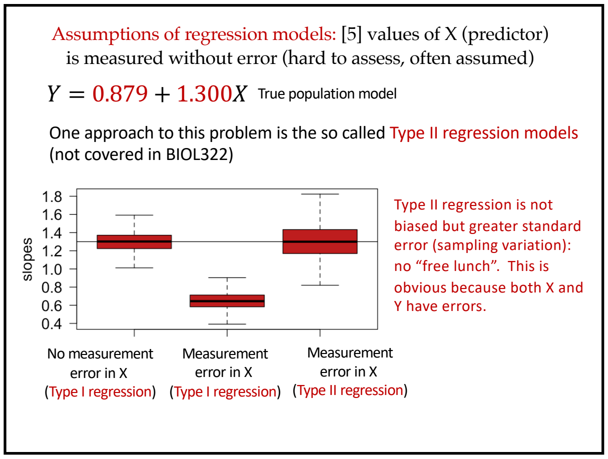
Assumptions of regression models: [5] values of X (predictor) is measured without error (hard to assess, often assumed)

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

43



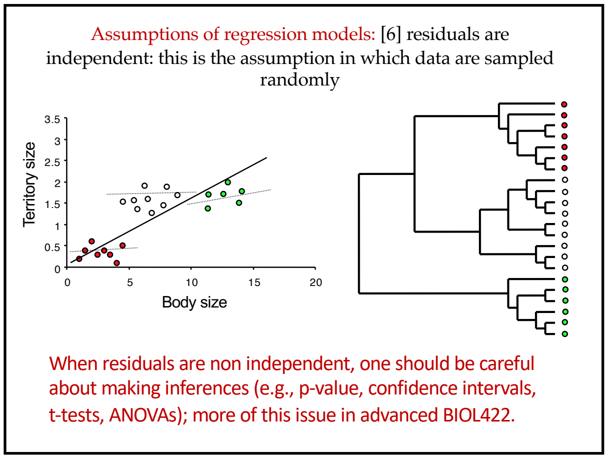
Assumptions of regression models: [5] values of X (predictor) is measured without error (hard to assess, often assumed)

$Y = 0.879 + 1.300X$ 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.

44



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.

45