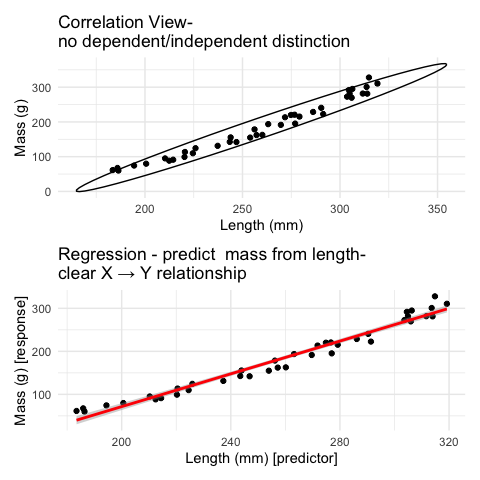
Lecture 10 - Regression and Linear Models

Your Name

# 1. Lecture 9: Review

Covered

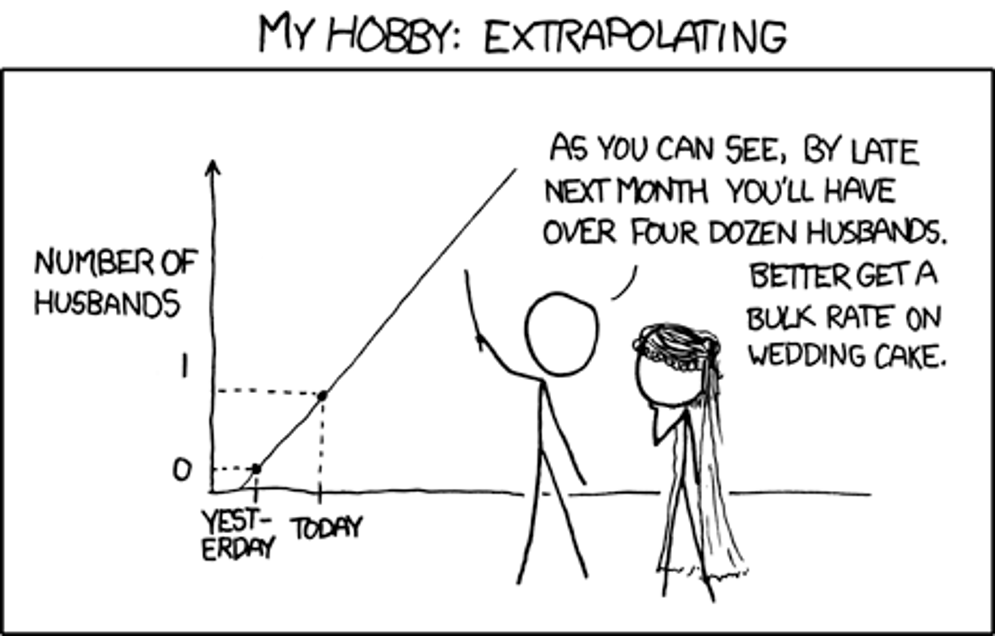
* Correlation analysis: measuring relationships between variables
* The distinction between correlation and regression
* Simple linear regression: predicting one variable from another
* Estimating and interpreting regression parameters
* Testing assumptions and handling violations
* Analysis of variance in regression
* Model selection and comparison



# 2. Lecture 10: Overview

Linear regression:

* REGRESSIONS:
  + Analysis of variance
  + Explained variance
  + Assumptions and diagnostics
  + Dealing w violations
  + Model II regression
  + Robust regression
* Smoothing Regressions



# 3. **Lecture 10:** Linear Regression

## 3.1 Simple Linear Regression Model

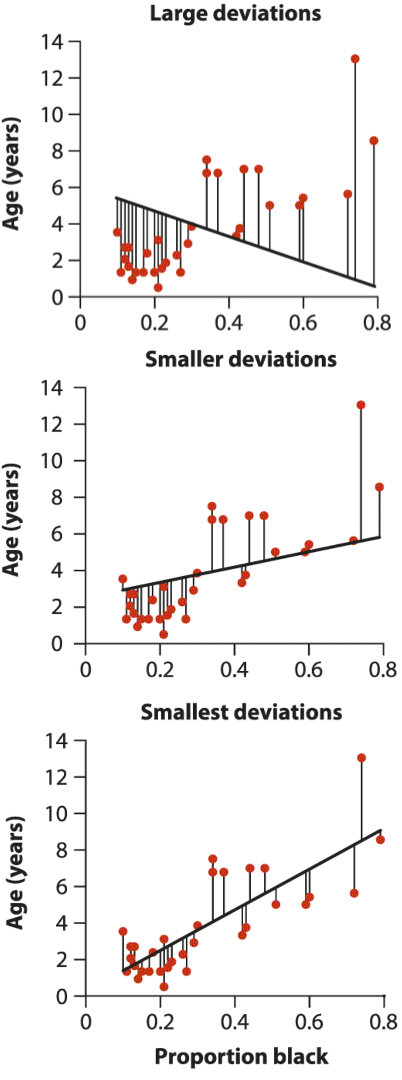
**Simple linear regression** models the relationship between a response variable (Y) and a predictor variable (X).

The **sample** regression equation is:

Where:

* is the predicted value of Y
* a is the estimate of α (intercept) sometimes
* b is the estimate of β (slope) sometimes

**Method of Least Squares**: The line is chosen to minimize the sum of squared vertical distances (residuals) between observed and predicted Y values.



# 4. **Lecture 10:** Linear Regression

## 4.1 Simple Linear Regression Model

**Simple linear regression** models the relationship between a response variable (Y) and a predictor variable (X).

The **sample** regression equation is:

Where:

* is the predicted value of Y
* a is the estimate of α (intercept) sometimes
* b is the estimate of β (slope) sometimes

**Method of Least Squares**: The line is chosen to minimize the sum of squared vertical distances (residuals) between observed and predicted Y values.



# 5. **Lecture 10:** Linear Regression

## 5.1 Simple Linear Regression Model

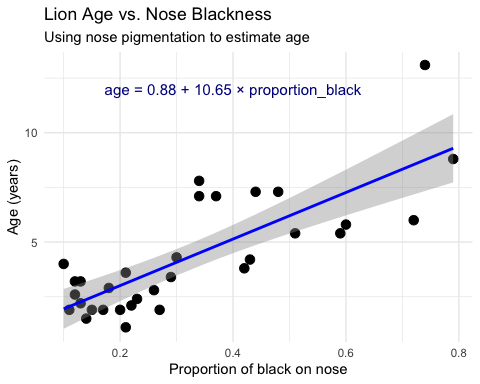
* male lions develop more black pigmentation on their noses as they age.
* can be used to estimate the age of lions in the field.

Call:  
lm(formula = age\_years ~ proportion\_black, data = lion\_data)  
  
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   
proportion\_black 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

# 6. **Lecture 10:** Linear Regression

## 6.1 Simple Linear Regression Model

* The calculation for slope (b) is:
* Given: -
* b = 13.0123 / 1.2221 = 10.647
* Intercept (a):
* **Making predictions:**
* To predict the age of a lion with 0.50 proportion of black on its nose:
* **Confidence intervals vs. Prediction intervals:**
  + **Confidence interval**: Range for the mean age of all lions with 0.50 black
  + **Prediction interval**: Range for an individual lion with 0.50 black
* Both intervals are narrowest near and widen as X moves away from the mean.



# 7. **Lecture 10:** Linear Regression - estimates of error and significance

In addition to getting estimates of population parameters (intercept - β0 , slope - β1)

want to test hypotheses about them

* This is accomplished by analysis of variance
* Partition variance in Y: due to variation in X, due to other things (error)



# 8. **Lecture 9:** Linear Regression - estimates of variance

Total variation in Y is “partitioned” into 3 components:

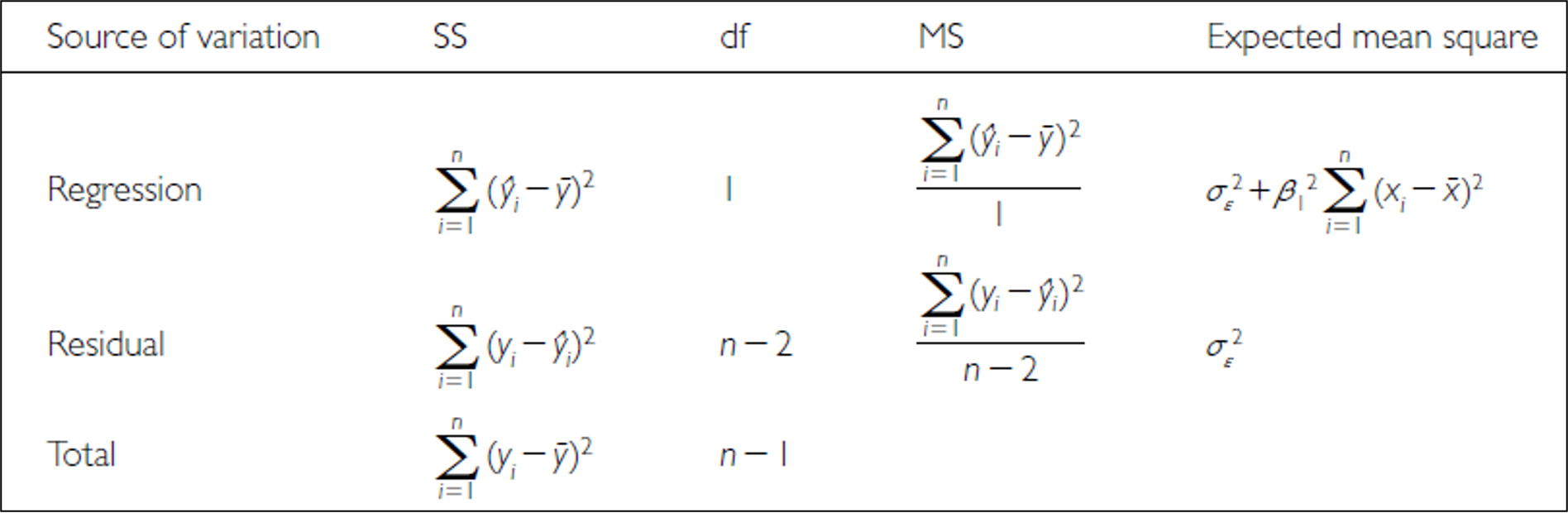
* : variation explained by regression
  + difference between predicted values (ŷi ) and mean y (ȳ)
  + dfs= 1 for simple linear (parameters-1)
* : variation not explained by regression
  + difference between observed () and predicted () values
  + dfs= n-2
* : total variation
  + sum of squared deviations of each observation () from mean ()
  + dfs = n-1



# 9. **Lecture 9:** Linear Regression - estimates of variance

Total variation in Y is “partitioned” into 3 components:

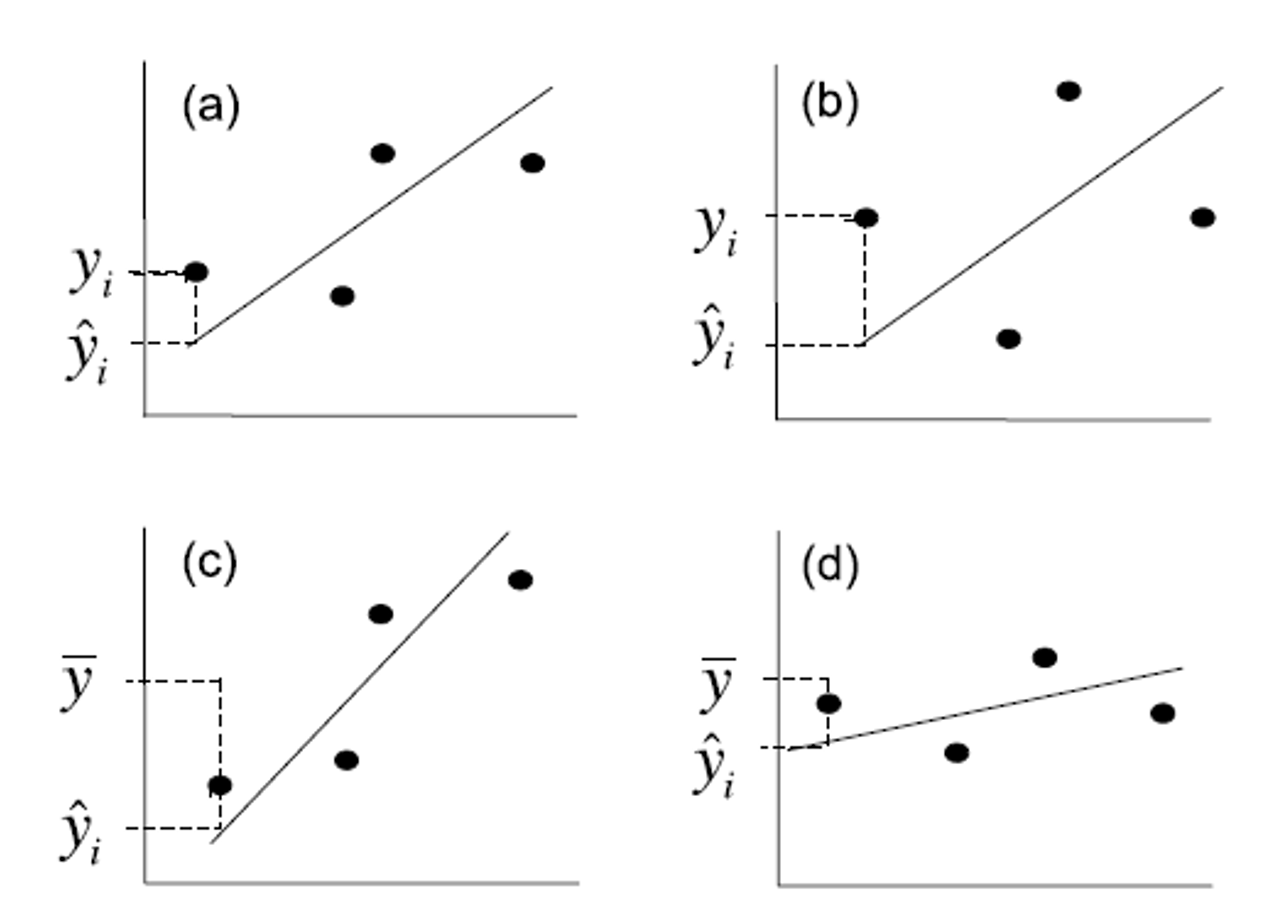
* : variation explained by regression
  + difference between predicted values (ŷi ) and mean y (ȳ)
  + dfs= 1 for simple linear (parameters-1)
* : variation not explained by regression
  + difference between observed () and predicted () values
  + dfs= n-2
* : total variation
  + sum of squared deviations of each observation () from mean ()
  + dfs = n-1



# 10. **Lecture 10:** Linear Regression - estimates of variance

Total variation in Y is “partitioned” into 3 components:

* : variation explained by regression
  + GREATER IN C than D
* : variation not explained by regression
  + GREATER IN B THAN A
* : total variation



# 11. **Lecture 10:** Linear Regression - estimates of variance

Sums of Squares and degress of freedome are:

* Sums of Squares depends on n
* We need a different estimate of variance



# 12. **Lecture 10:** Linear Regression - estimates of variance

Sums of Squares converted to Mean Squares

* Sums of Squares divided by degrees of freedom - does not depend on n
* : estimate population variation
* : estimate pop variation and variation due to X-Y relationship
* Mean Squares are not additive



# 13. **Lecture 10:** Linear Regression - Null Hypothesis

Regression typically tests null hypothesis that β1 = 0

* or no relationship between X and Y

Can test in two ways:

## 13.1 Using t-statistic:

* = Standard error of slope estimate
* Bo= 0: t-test:
* 1 parameter t-test, where testing whether β1 =0
* t-statistic test is more general
* R can provide both
* Can also ask whether β0 =0 using t-test
* or whether two regression lines are significantly different

# 14. **Lecture 10:** Linear Regression - Null Hypothesis

Regression typically tests null hypothesis that β1 = 0

* or no relationship between X and Y

Can test in two ways:

## 14.1 Using F-ratio:

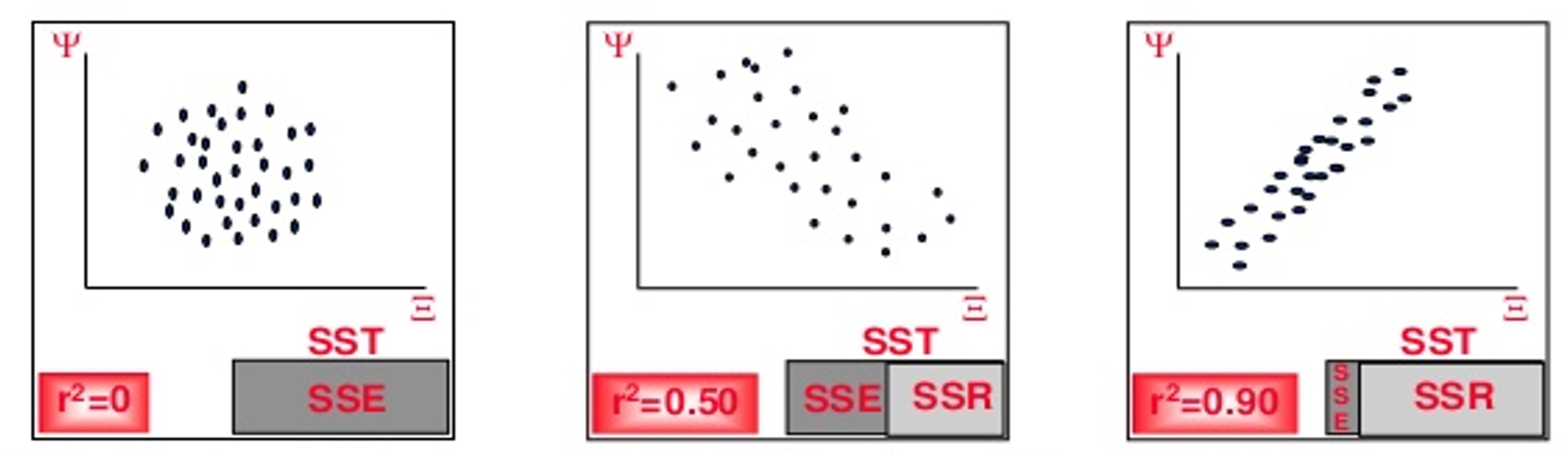
* if β1 = 0, ratio will be = 1 otherwise >1
* compare F-ratio to df-specific F-distribution
* decide how likely obtain our F-ratio by chance

# 15. **Lecture 10:** Linear Regression - Explained variance

* Want to know how strong is association between X and Y
* Coefficient of determination (): proportion of variation in Y explained by X
* When more of variation is due to regression rather than ‘error’, closer to 1

# 16. **Lecture 10:** Linear Regression - Explained variance

* Want to know how strong is association between X and Y
* Coefficient of determination (): proportion of variation in Y explained by X
* When more of variation is due to regression rather than ‘error’, closer to 1



# 17. **Lecture 10:** Linear Regression - Explained variance

summary(lion\_model)

Call:  
lm(formula = age\_years ~ proportion\_black, data = lion\_data)  
  
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   
proportion\_black 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

anova(lion\_model)

Analysis of Variance Table  
  
Response: age\_years  
 Df Sum Sq Mean Sq F value Pr(>F)   
proportion\_black 1 138.544 138.544 49.75 7.677e-08 \*\*\*  
Residuals 30 83.543 2.785   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 18. Reporting results

“Lion age (years) could be predicted from nose spots (percentage) using the simple linear regression model age = 10.67 \* proportion\_black + 0.8749. Regression analysis showed that the slope of the relationship was significantly (at α=0.05) different than 0 ( = 49.93, p < 0.0001, R² = 0.6247).”

Note there is an adjusted R² - what is that - accounts for the number of predictors in your model - adjusts the R² value by penalizing the addition of variables that don’t improve the model fit significantly

The formula for adjusted R² is:

Where:

* n is the number of observations (32 lions)
* p is the number of predictors (1 = proportion\_black)

measures the proportion of variance in the dependent variable (age\_years) that is explained by the independent variable (proportion\_black).

# 19. Assumptions and diagnostics of regression

* Assumptions apply to observed values of Y and εi
* most can be assessed by looking at residuals (distance from predicted)

**Linearity:**

* relationship between X and Y in population is straight line

**Check:**

* examine biplot of Y on X

**If violated:**

* transform Y
* use polynomial or nonlinear regression



# 20. Assumptions and diagnostics of regression

**Normality:**

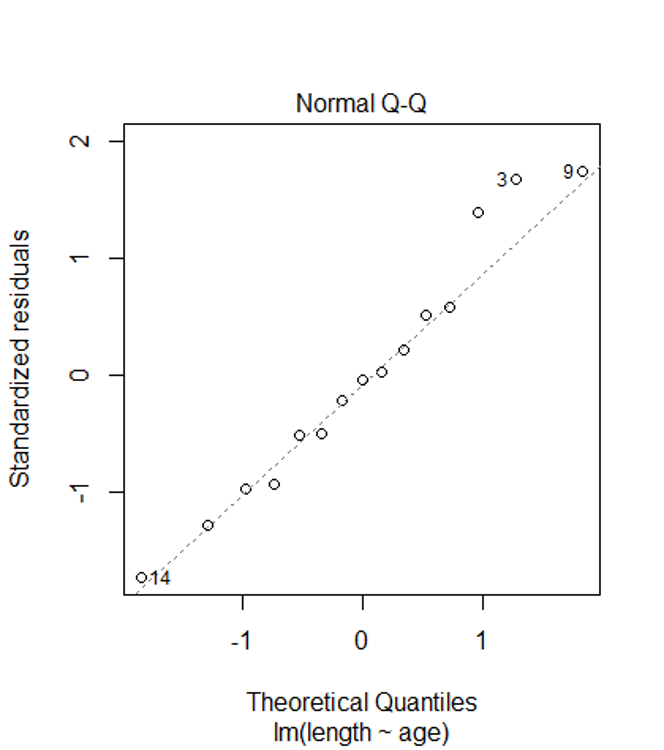
* y-values for each xi are normally distributed.
* OLS estimates moderately robust to violation

**Check:**

* are residuals normally distributed?
* qq plots, histogram of residuals, shapiro-wilk test

**If violated:**

* transform Y
* use Generalized Linear Model



# 21. Assumptions and diagnostics of regression

**Homogeneity of variance:**

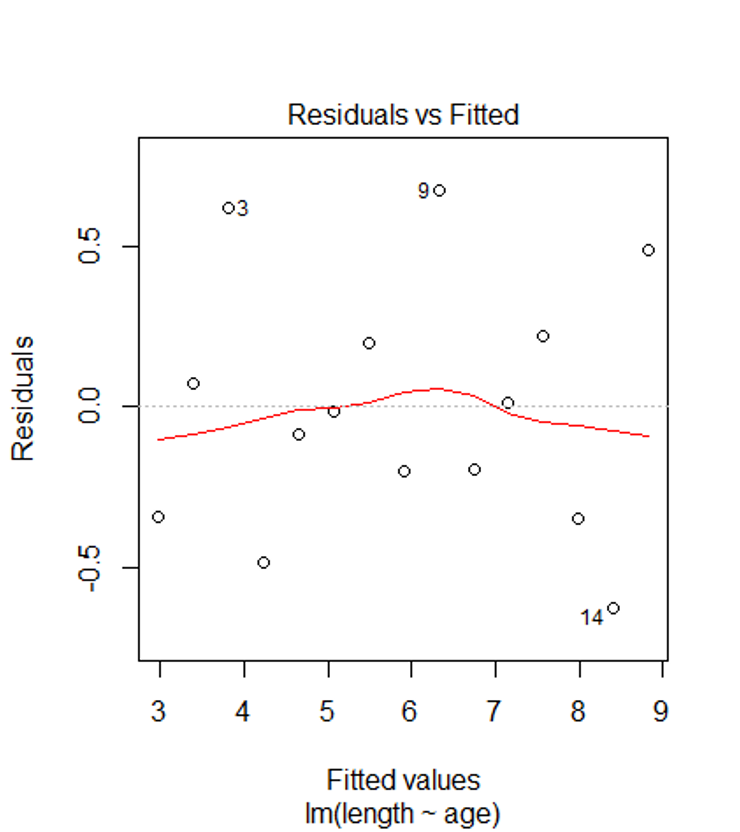
* y-values for each xi have same variance.
* OLS estimates *NOT* robust to violation

**Check:**

* plot residuals against x-values or predicted values (ŷi)

**If violated:**

* transform Y
* use GLM
* weighted LS regression



# 22. Assumptions and diagnostics of regression

**Independence**:

* Y values from each xi do not influence each other
* Often violated with repeated measurements in time/space -> autocorrelation

**Check**: determine correlation coefficient bw adjacent residuals

**If violate**:

* ANOVA (grouping present)
* mixed model ANOVA
* time series



# 23. Assumptions and diagnostics of regression

**Fixed X:**

* xi are known values fixed by researcher (e.g., drug doses).
* Often not true in ecology.

**If violated**:

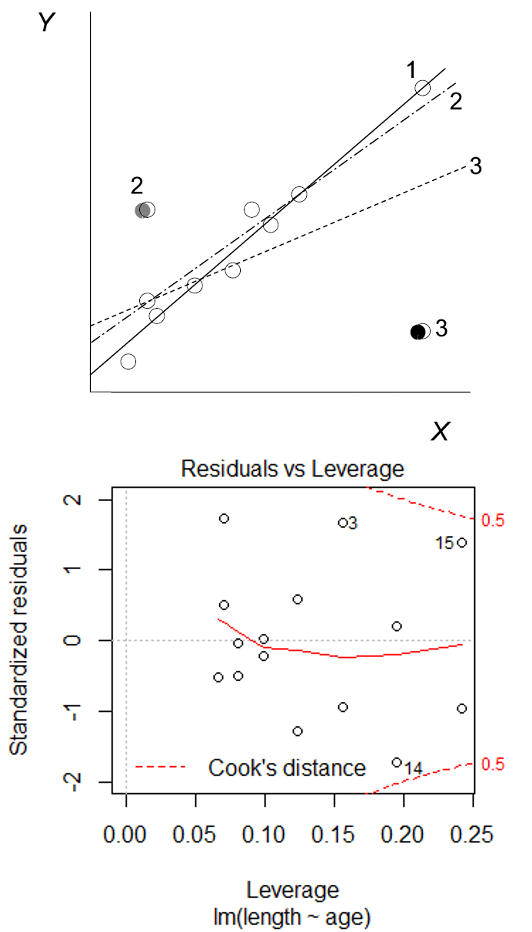
* not problem for H-testing, prediction, **but**
* error underestimated.
* Can use model II regression



# 24. Assumptions and diagnostics of regression

Outlier or Influence:

* how much each point affects slope (Cook’s D)
* Large Di (>1) indicates influential observation

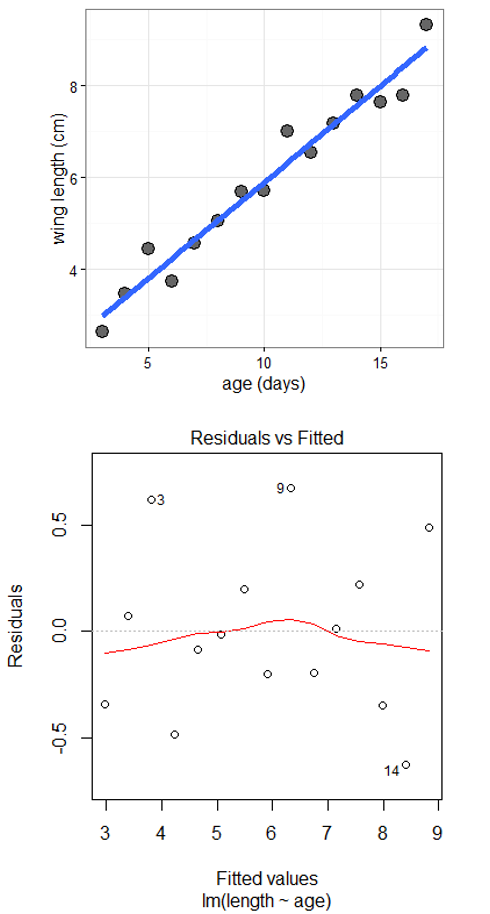


# 25. Assumptions and diagnostics of regression

**Residual plots**

residuals vs. predicted y:

* can be used to assess assumptions:
  + linearity
  + normality
  + equal variance
  + outliers



# 26. Dealing with violations

Weighted least squares:

* when variance unequal, can use WLS approach
* each point weighted by reciprocal of variance (points w large variance given less weight)

Robust regression:

* when distribution distinctly non-normal and/or large outliers

LAD:

* parameters estimated from non-squared residuals
* outliers not as influential

M-estimators:

* residuals have different weight depending on distance from mean

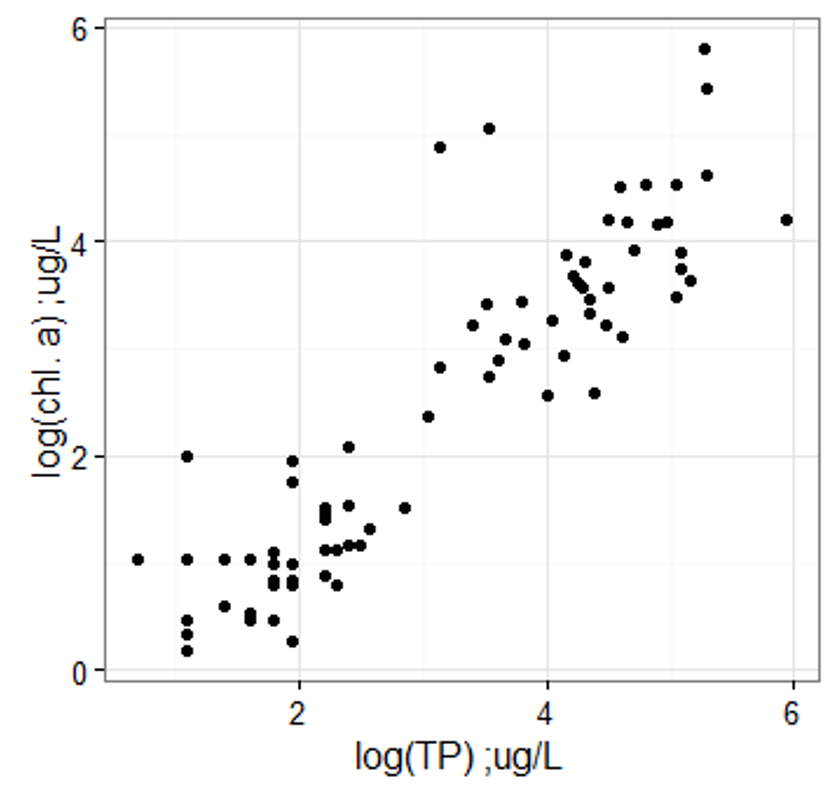
Rank-based: “if all else fails”

# 27. Model II regression

Fixed X is assumption of regular regression

what if X random (typical case)?

* If goal is prediction (interpolation) then Model I is ok…
* if goal is correct parameters and error estimates, may need to use Model II

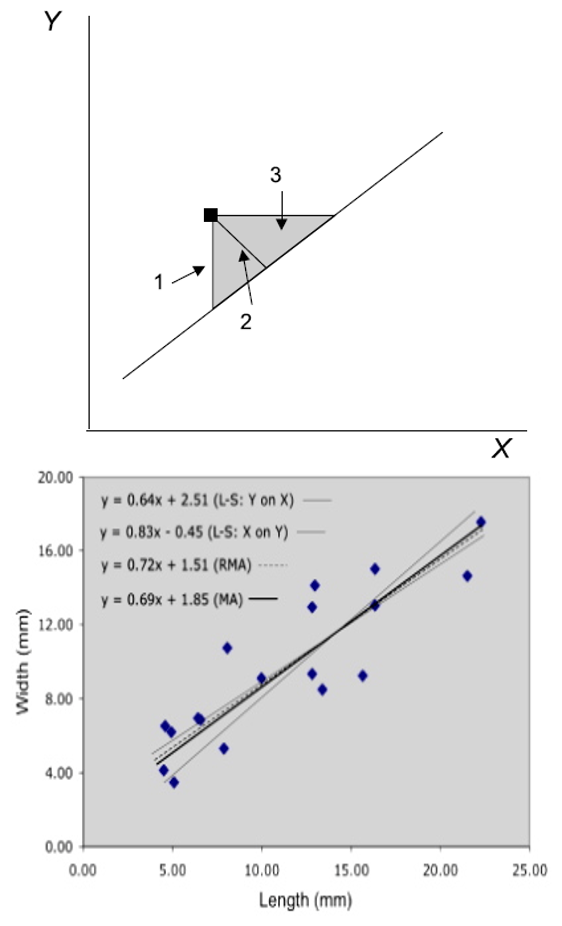


# 28. Model II regression

Model II regression - Approach underused in ecology

* Model I regression will still perform well for H-tests
* underestimate true slope
* minimizes distance between points and line along both axes
* (vs. on Y only in OLS)

MA and RMA approach slightly different



# 29. Model II Regression

## 29.1 Detailed Explanation of Model II Regression Types

1. Standardized Major Axis (SMA)

* SMA regression minimizes the product of the vertical and horizontal distances from the points to the regression line. It’s implemented in the smatr package with method=“SMA”. SMA is appropriate when the measurement scales of X and Y are different.

1. Major Axis (MA)

* MA regression minimizes the perpendicular distances from the data points to the regression line. It’s implemented in the smatr package with method=“MA”. MA is appropriate when X and Y are measured in the same units.

1. Reduced Major Axis (RMA)

* RMA regression (also called geometric mean regression) is available in the lmodel2 package. It produces a slope that is the geometric mean of the OLS regression slopes of Y on X and X on Y (specifically, it equals the OLS slope of Y on X multiplied by the sign of the correlation between X and Y, divided by the square root of the R² value).

# 30. Model II Regression

## 30.1 When to Use Each Method

OLS (Model I) - Use when:

* X is measured without error
* The research goal is predicting Y from X
* There’s a clear dependent variable

MA (Major Axis) - Use when:

* X and Y are measured in the same units
* Both variables have similar error variances
* The goal is to understand the symmetric relationship

SMA (Standardized Major Axis) - Use when:

* X and Y are measured in different units
* The goal is to understand the structural relationship
* You want to test for isometry or allometry in scaling studies

RMA (Reduced Major Axis) - Use when:

* The ratio of error variances is approximately equal to the ratio of the true variances
* Both variables contain measurement error
* Neither variable is clearly dependent or independent

# 31. Model II Regression

## 31.1 Key Differences in Results

The slopes of these methods will typically follow this pattern when the correlation coefficient is less than 1: OLS slope < MA slope < RMA slope < inverse of OLS (X on Y) slope This is particularly evident when the correlation between X and Y is weaker. As correlation approaches 1, the differences between methods diminish.

# 32. Model II Regression

## 32.1 Decision Tree

Here’s a simplified decision tree:

-Are X and Y measured with error? If No → Use OLS (Model I) -Are the errors in X and Y approximately equal? If Yes → Use MA -Are X and Y measured in different units/scales? If Yes → Consider SMA -Is the correlation between X and Y weak (<0.7)? If Yes → Method choice is critical; consider RMA -Are you uncertain about error structure? If Yes → RMA is a reasonable compromise

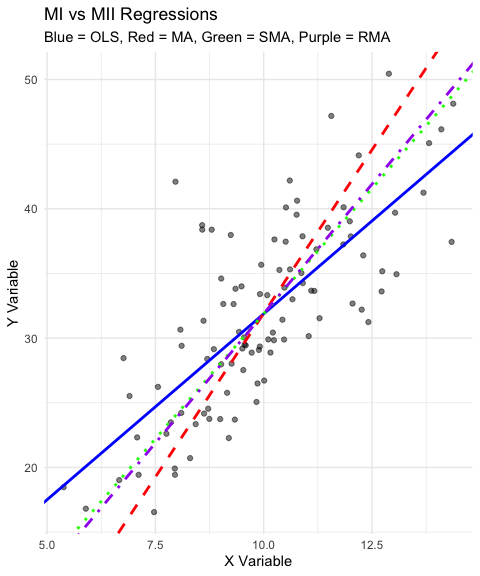
Remember that when the correlation between X and Y is very strong (r > 0.9), all methods will yield similar results, making the choice less critical. The differences between methods become more pronounced as the correlation weakens.

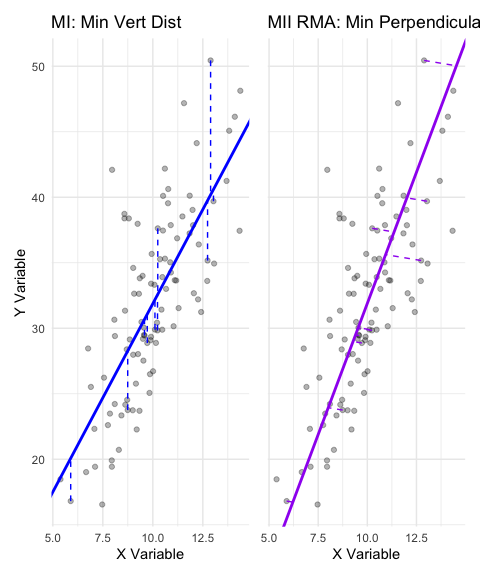
Finally, it’s often valuable to run multiple methods and compare the results. If they lead to different ecological or biological interpretations, this should be explicitly addressed in your discussion.

# 33. Model II Regression

Call:  
lm(formula = y ~ x, data = data\_ols\_m2)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-8.058 -3.498 -0.990 2.946 16.070   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 3.1346 2.6584 1.179 0.241   
x 2.8745 0.2617 10.982 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.858 on 98 degrees of freedom  
Multiple R-squared: 0.5517, Adjusted R-squared: 0.5471   
F-statistic: 120.6 on 1 and 98 DF, p-value: < 2.2e-16

Model II regression  
  
Call: lmodel2(formula = y ~ x, data = data\_ols\_m2, range.y =  
"relative", range.x = "relative", nperm = 99)  
  
n = 100 r = 0.7427668 r-square = 0.5517025   
Parametric P-values: 2-tailed = 9.070988e-19 1-tailed = 4.535494e-19   
Angle between the two OLS regression lines = 8.317517 degrees  
  
Permutation tests of OLS, MA, RMA slopes: 1-tailed, tail corresponding to sign  
A permutation test of r is equivalent to a permutation test of the OLS slope  
P-perm for SMA = NA because the SMA slope cannot be tested  
  
Regression results  
 Method Intercept Slope Angle (degrees) P-perm (1-tailed)  
1 OLS 3.134600 2.874473 70.81773 0.01  
2 MA -18.688419 5.059928 78.82062 0.01  
3 SMA -6.805843 3.869953 75.51163 NA  
4 RMA -8.131038 4.002664 75.97273 0.01  
  
Confidence intervals  
 Method 2.5%-Intercept 97.5%-Intercept 2.5%-Slope 97.5%-Slope  
1 OLS -2.140952 8.410152 2.355051 3.393894  
2 MA -29.748586 -10.906922 4.280654 6.167542  
3 SMA -12.339088 -1.965648 3.385234 4.424077  
4 RMA -16.211541 -1.554125 3.344023 4.811882  
  
Eigenvalues: 54.08817 1.502864   
  
H statistic used for computing C.I. of MA: 0.001181286





2.5 % 97.5 %  
(Intercept) -2.140952 8.410152  
x 2.355051 3.393894

Call: sma(formula = y ~ x, data = data\_ols\_m2, method = "SMA")   
  
Fit using Standardized Major Axis   
  
------------------------------------------------------------  
Coefficients:  
 elevation slope  
estimate -6.805843 3.869953  
lower limit -12.094381 3.385234  
upper limit -1.517306 4.424077  
  
H0 : variables uncorrelated  
R-squared : 0.5517025   
P-value : < 2.22e-16

Call: sma(formula = y ~ x, data = data\_ols\_m2, method = "MA")   
  
Fit using Major Axis   
  
------------------------------------------------------------  
Coefficients:  
 elevation slope  
estimate -18.688419 5.059928  
lower limit -27.905282 4.280260  
upper limit -9.471556 6.168337  
  
H0 : variables uncorrelated  
R-squared : 0.5517025   
P-value : < 2.22e-16